

Waking up to the importance of sleep

A growing chasm separates the growing scientific understanding of sleep, and the widespread public assumption that it just doesn't matter.

Is a full night's sleep an essential ingredient for a healthy mind and body, or just a luxury for the lazy? Despite extensive evidence for the former, the tendency of society (including hypercompetitive heads of laboratories) to applaud people who claim to be forever on the go leaves many assuming the latter. The notion that successful people can get by with just a few hours of sleep a night reinforces a common perception that sleep is a waste of time.

This misperception carries serious ramifications. The tendency to sleep less — perhaps 20% less in industrialized countries than a century ago — has serious consequences for public safety. In a poll funded by the US National Science Foundation (NSF), more than one-third of American drivers admit to having nodded off behind the wheel. Public disasters such as the 1989 *Exxon Valdez* oil-tanker accident have been attributable, in large part, to sleep deprivation. Furthermore, a mounting body of evidence links a lack of sleep to mood swings, depression, anxiety and other mental illness.

The medical profession has been slow to acknowledge all this. According to the same NSF poll (www.sleepfoundation.org), 75% of patients report sleep problems, but less than one-third say their doctor has asked them about it. And trainee physicians seem to be trained to wear their own sleep deprivation as a badge of honour, rather than what it is — a threat to health, safety and professionalism.

A *Nature* Insight on sleep in this issue (see page 1253) sheds some fascinating light on sleep research. It explores questions ranging from what we might learn about the function of sleep by studying the variation in sleep patterns between different mammals, to whether sleep is essential for some forms of memory consolidation.

Sleep is much more than an absence of activity in the body and brain. The brain is highly active during sleep, especially during REM sleep, which is characterized by rapid eye movements and vivid dreaming. Sleep is also far from the single phenomenon it is sometimes assumed to be: the brain activities behind its different stages can be as distinct from each other as they are from wakefulness.

Although answers to some of the most basic questions in sleep research — such as why we sleep at all — remain elusive, recent

developments do provide knowledge that is directly relevant to public health and safety. Despite the common perception that falling asleep is a gradual process, for example, the transition from being awake to being asleep can be extremely rapid.

On page 1257 of this issue, Harvard neuroscientist Clifford Saper and his colleagues describe how a 'flip-flop' switch in the part of the brain known as the hypothalamus brings about discrete transitions between waking and sleeping. They show how damage to this switching mechanism can lead to instabilities in both sleeping and waking states, as seen in narcolepsy, a debilitating neurological disease. The very concept of a switch that seems optimized to flip rapidly between being asleep and awake, without an intermediate state, should give pause to anyone tempted to drive or operate machinery when feeling drowsy.

Researchers should, however, be cautious about overplaying data that link sleep with illness. Both heart disease and obesity, for example, have been linked to sleep problems in some studies. But despite clear signs of a correlation, evidence of a causal link remains inconclusive.

It would be naive to think that simply laying out the scientific facts will change popular perceptions about sleep and behaviour. Data proving the link between smoking and lung cancer, for example, were available long before behaviour changed on a significant scale. Even with the assistance of extensive public information campaigns, the message took a long time to sink in.

Nonetheless, basic neuroscience research has a role not just in building our understanding of sleep, but also in helping to convince the public to take it seriously. Public communication of important findings should be helped by the fact that so many people find the topic fascinating. Basic researchers, clinicians and educators can take advantage of this interest to make the case for changes in behaviour — while ensuring that they, too, get a good night's sleep. ■

"Neuroscience research has a role not just in building our understanding of sleep, but also in helping to convince the public to take it seriously."

Is the city safe?

The Environmental Protection Agency is ducking a frank assessment of New Orleans after Katrina.

After the terrorist attacks of 11 September 2001, the US Environmental Protection Agency (EPA) took some flak for declaring, perhaps prematurely, that the air was safe in the vicinity of the World Trade Center in New York.

Now the agency, which is charged with protecting public health

and the environment, is abdicating its responsibility to issue clear public guidance on possible health hazards in New Orleans, flooded last month by Hurricane Katrina. It is saying nothing on the advisability of returning to the ruined city, arguing instead that its job is just to run tests and pass on data to local officials, who will make of them what they will.

For what was once the world's foremost environmental agency, this simply isn't good enough. The EPA's own scientific advisory board, as well as the usual welter of environmental groups, are rightly calling on the agency to do its job properly, and give the American people more solid information about the environmental



risks posed by episodes such as the flooding after Hurricane Katrina (see page 1216).

Members of the science advisory board are unhappy with the tests that were carried out in the first days, when the city was still under water. Tests were done for many regulated chemicals, but those for more obvious threats, such as disease-causing microbes, were not. According to Granger Morgan, a technology policy expert at Carnegie Mellon University in Pittsburgh and chairman of the EPA science advisory board, the agency needs to prepare better plans for specific emergency situations, so that it can respond appropriately.

Most of the data currently being collected by the agency on the ground in New Orleans pertain only to short-term risks. Levels of metals or pesticides in the sediment left behind as the floodwater receded are being compared with exposures that are safe for a few days, or even less. People who want to find out whether the levels of contaminants near their homes are dangerous in the longer term will have to do their own research.

The agency can hardly be accused of sitting around twiddling its thumbs in New Orleans. It has more than 1,000 employees working in the ravaged track of the storm, and is doing its best to advise members of public about how best to protect themselves from contaminants. Hundreds of measurements have been carefully posted on its website (www.epa.gov/katrina). All of this has been done

within the agency's existing and rather overstretched budget.

People are moving back into New Orleans now, pushing their way into mud-caked buildings, sleeping in rotting, oily houses, and scrubbing mould off the walls without wearing protection — or, in at least one case, with respirators gamely strapped on upside-down.

Naturally, the political pressure to repopulate the area is intense. There is nothing to suggest that the city should be declared uninhabitable. But the public deserves much more than statements such as one issued on 17 September, to the effect that neither the EPA nor the Centers for Disease Control and Prevention in Atlanta, Georgia, will come forward to offer any guidance on the reinhabitation of New Orleans.

As the EPA's own inspector-general declared in 2003: "EPA needs to be prepared to assert its opinion and judgment on matters that impact human health and the environment, regardless of who else is involved or may share responsibility. Ultimately, the public, Congress, and others expect EPA to monitor and resolve environmental issues." In the wake of Katrina, the need for the agency to fulfil that role is clearer than ever. ■

"EPA needs to be prepared to assert its opinion and judgment on matters that impact human health and the environment, regardless of who else is involved."

Free tips

The National Academies offers guidance to keep the United States internationally competitive.

Worthy reports are as frequently encountered in Washington DC as grand monuments — and are just as likely to be ignored by the locals. But lawmakers will fail their constituents if they manage to ignore the latest study on competitiveness by the National Academies.

The report, *Rising Above the Gathering Storm*, addresses resurgent fears that the United States' longstanding global leadership in research and development is on the wane. Written by a panel chaired by Norman Augustine, former chief executive of technology corporation Lockheed Martin, it offers several concrete recommendations designed to keep that leadership intact. Whether the US government pays any attention or not, its competitors will find the panel's findings well worth a look.

Whereas previous studies of this kind have focused primarily on research funding, this one concentrates much of its attention on improving the nation's scientific literacy. It calls for the annual recruitment of some 10,000 science and maths teachers, proposing that science undergraduates be lured into the classroom with generous scholarships, with the lofty goal of improving science education at school for some 10 million people.

This proposal may appear to some US politicians to be central planning run amok — and it doesn't really address the low pay and social standing of teachers in the United States. But it does have potential and precedent: an existing programme called Teach for America has succeeded in recruiting thousands of young college

graduates to teach in the nation's most troubled neighbourhoods.

Other recommendations of note include a call for the creation of a new energy-research agency that would conduct low-cost, high-risk, high-reward research projects. This would be modelled on the Defense Advanced Research Projects Agency (DARPA), which has been highly successful in backing exciting, basic research that may spawn useful technology. Another new entity would be set up expressly to arrange for the construction of scientific facilities.

The academy panel also proposes radical changes in the treatment of young scientists in general, and foreign ones in particular. It suggests a new category of generous grants that would allow young researchers early in their career to firmly establish their own lines of enquiry. Furthermore, it calls for changes in US immigration policy that would make it easier for foreign students and scientists to stay in the country to continue their careers. Both these suggestions will require serious political commitment to implement — but they would go a long way towards fostering fresh scientific talent.

Senators Lamar Alexander (Republican, Tennessee) and Jeff Bingaman (Democrat, New Mexico), who commissioned the study, must now try to drum up support on Capitol Hill for the implementation of its recommendations. They face an uphill battle, given the size of the US budget deficit and inevitable political resistance to such concepts as further federal involvement in school education. The United States' competitors, in Europe and the Far East, also need to consider such measures, and might actually find some of them easier to implement. ■

"Some of these suggestions will require serious political commitment to implement — but they would go a long way towards fostering fresh scientific talent."

RESEARCH HIGHLIGHTS

EVOLUTION

Fickle enzymes

Science **310**, 499–501 (2005)

It is surprisingly easy for enzymes to switch their preference for the molecule they work with, according to US researchers. Their study shows that a small number of mutations can dramatically alter a protein's function. This allows the protein to evolve new abilities quickly, without taking intermediate forms that harm the organism's fitness.

Antony Dean and his team at the University of Minnesota, St Paul, studied an enzyme called isopropylmalate dehydrogenase, or IMDH. IMDH needs the help of another molecule, called NAD, to function. Working in bacteria, Dean's team altered amino acids in IMDH and found that just five swaps were needed to make IMDH switch to a different helper molecule known as NADP.

MOLECULAR BIOLOGY

Double defence

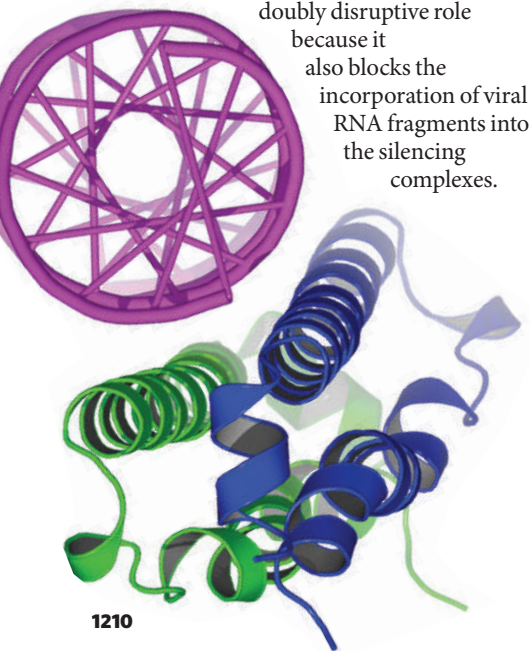
Nature Struct. Mol. Biol. doi:10.1038/nsmb1005 (2005)

The structure of a protein (pictured below) produced by the insect- and plant-infecting Flock House virus sheds light on how some viruses might counter their hosts' defences.

Two B2 proteins (blue and green) form a four-helix bundle that binds to the virus's double-stranded RNA (pink). This protects the viral RNA against cleavage — the host cell needs to cleave viral RNA to build RNA silencing complexes in response to infection.

The researchers, led by James Williamson of the Scripps Research Institute in La Jolla, California, suggest that B2 has a doubly disruptive role

because it also blocks the incorporation of viral RNA fragments into the silencing complexes.

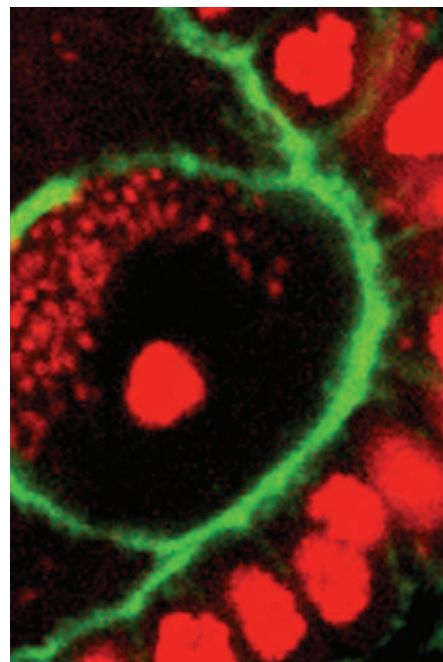


From mother with love

PLoS Pathogens **1**, e14 (2005)

Wolbachia — a parasitic bacterium that lives inside the cells of many insect species — is transmitted by females to their offspring. William Sullivan from the University of California, Santa Cruz, and his colleagues have looked in unprecedented detail at this process.

The parasite resides in the cytoplasm, a component of eggs but not of sperm. The researchers showed that, in the fruitfly *Drosophila*, *Wolbachia* infects immature egg cells and multiplies rapidly during the egg's development. They also found that the bacteria cluster at the top of the egg (pictured; *Wolbachia* and host DNA shown in red). Both the localization and rapid multiplication depend on *Wolbachia* commandeering its host's transport system, consisting of the cell's network of microtubules and associated motor proteins.



P. M. FRYDMAN & W. SULLIVAN

NANOTECHNOLOGY

Live wire

Angew. Chem. Int. Edn **44**, 2–7 (2005)

In an exciting union of microbe and machine, living bacteria have been incorporated into an electronic circuit to produce a humidity sensor.

Vikas Berry and Ravi Saraf from the University of Nebraska, Lincoln, placed *Bacillus cereus* bacteria on a silicon chip inlaid with gold electrodes, then applied a wash containing gold particles measuring just 30 nanometres across. This covered the microbes with a bristle-like gilt that conducts electricity.

The size of the electrical current depends on the separation of the particles, so a rise in moisture levels that causes the bacteria to swell is detected as a drop in current. Tests indicate that the device is much more sensitive than a conventional humidity gauge.

CANCER

Two roles in tumours

Nature Cell Biol. doi: 10.1038/ncb1314 (2005)

It's easy to tire of learning about players in the p53 pathway, but this research reveals an interesting case. It explains how the transcription factor KLF4 can suppress tumour growth in some situations, yet act as a tumour promoter in others.

Daniel Peeper and his colleagues report that KLF4 regulates, in opposing directions, two genes central to growth control. It suppresses p53, whose gene product

suppresses tumours, and it induces p21CIP1, a gene controlling normal cell proliferation.

The researchers, from the Netherlands Cancer Institute in Amsterdam, demonstrate that KLF4 becomes a growth promoter when p21CIP1 is dysfunctional — as occurs in most cancers. In these genetic circumstances, KLF4's suppression of p53 goes unchecked.

ASTRONOMY

X-ray vision

Astrophys. J. **632**, L99–L102 (2005)

The discovery of X-rays coming from a 35-year-old supernova, 1970G, has helped astronomers to extract data on the object from older, lower resolution observations.

Compiling these data gives a unique record that traces the object's X-ray emissions from when the star exploded through its transition to a supernova remnant. The X-rays probably emanate from material heated up by a shock wave closing in on the remnant's centre. The X-rays were discovered using a powerful space telescope — the Chandra X-Ray Observatory — and were analysed by Stefan Immler and Kip Kuntz of the NASA Goddard Space Flight Center in Greenbelt, Maryland.

NEUROSCIENCE

Thumbs up

J. Neurosci. **25**, 9339–9346 (2005)

Just observing an action can lead to the formation of a corresponding motor memory, according to a study led by Joseph

NATURE STRUCT. MOL. BIOL.

Classen at the University of Würzburg in Germany, and Leonardo Cohen of the National Institute of Neurological Disorders and Stroke, based in Bethesda, Maryland.

The researchers applied a magnetic field to the primary motor cortex region of the brain in human subjects, provoking a thumb twitch that was biased in one direction. The subjects then watched thumbs being twitched in a different direction. When the magnetic field was applied again, in the same way as before, their thumb movements were more likely to match the second, observed action.

The researchers say the formation of the memory may be linked to mirror neurons in the primary motor cortex, which fire both when an action is performed and when it is observed.

MATERIALS

Evidence of distortion

Science **310**, 468–470 (2005)

A debate over the complex behaviour of solids made from carbon buckyballs may be settled by recent experimental evidence.

Using a scanning tunnelling microscope, researchers headed by Michael Crommie from the University of California, Berkeley, found that deforming the football-shaped spheres of 60 carbon atoms helped to make the material switch from behaving as a metal, to behaving as an insulator.

Although the idea has been mooted before, this study presents images of a charge-induced Jahn–Teller distortion occurring in a monolayer of the material deposited on gold. The layer's properties change as it is doped with potassium atoms, which transfer electrons to the spheres. The build-up of charge leads to the carbon cages being squished across one axis, the team reports.

OPTICS

Bright sparks

Phys. Rev. Lett. **95**, 143902 (2005)

A 'molecule' of light has been sent down an optical fibre by Martin Stratmann and his colleagues at the University of Rostock in Germany.

The molecule was made from two optical solitons, which are pulses of light that maintain their shape and intensity. The researchers showed that the solitons could be bound to one another, so that a pair can, in principle, travel tens of kilometres down optical fibres (pictured right) without becoming separated — in the same way that a diatomic molecule might drift through space. They suggest the robust doublet could be used in optical data transmission to represent a third character, in addition to the binary 0 (dark) and 1 (a single bright pulse).

NEUROBIOLOGY

Calling time

Neuron **48**, 213–219; 221–227; 267–278 (2005)

Starting from entirely different points, three groups have identified the receptor for a neuropeptide that helps to regulate circadian rhythms in insects. The receptor, which is for pigment-dispersing factor (PDF) in the fruitfly *Drosophila*, is a class II G-protein-coupled receptor dubbed CG13758. It appears in neurons that make up the master clock, and in some outside this time-keeping centre. Moreover, genetic manipulation of the receptor's expression suggests that PDF may be part of the network that synchronizes the rhythms of individual clock neurons.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

CELL BIOLOGY

RNA in reserve

Cell **123**, 249–263 (2005)

Researchers believe they may have discovered a new class of messenger RNA.

David Spector of Cold Spring Harbor Laboratory, New York, and his co-workers have found that some RNA transcripts of a mouse gene called *mCAT2* remain as back-up copies in the cell nucleus, only being released into the cytoplasm when the cell is stressed.

The normal protein product of *mCAT2* helps cells to respond to pressures such as viral infection; so mobilizing this reserve RNA might allow the cell to react more quickly. The researchers have named the reserve strands CTN-RNA, for CAT2-transcribed nuclear RNA.

JOURNAL CLUB

Monica Justice
Baylor College of Medicine,
Houston, Texas

A researcher from the mouse genome project argues that comparing sequences could help to explain gene regulation.

"Bioinformatics is not science," grumbled the student, frustrated with our committee's insistence that he compare his bat sequence to several other genomes.

To us, the sequence was

incomplete without such analysis. To the student, it was a tedious exercise that would yield no mechanistic insight.

But it might well open other doors. I am reminded of how the discovery of the four bases of DNA led to a crucial understanding of how proteins are encoded. Similarly, learning how genomes are related and organized over evolutionary time may teach us more about how chromosomes and genomes work.

In trying to make sense of genomic data, Adam Siepel from

the University of California, Santa Cruz, and his colleagues have realized that to compare the genomes of highly divergent species, sequences need to be calibrated.

Millions of years after species split, non-coding regions of DNA are expected to contain more base substitutions than coding regions. Siepel's team came up with a model to predict this variation, then adapted a computer program that compares genomes to correct for it. Searches among five vertebrate, four insect, two worm

and seven yeast species revealed some surprisingly long, highly conserved elements (Siepel, A. *et al. Genome Res.* **15**, 1034; 2005).

New ways of looking at the big picture are bound to lead to further discoveries. There are hints that the elements that Siepel's group have found play a role in regulating genes, through RNA editing and splicing. If we help the student to do his comparison, he might just find that the difference between the bat wing and the human hand is down to regulation too.

NEWS

Migration threatens to send flu south

Researchers fear that the bird flu virus's next stop will be Africa, where dependence on poultry means that the consequences could be even worse than in southeast Asia. **Tom Simonite** reports.

The deadly H5N1 bird flu virus is expected to be carried by birds into the Middle East and east Africa within weeks, the United Nations' Food and Agriculture Organization (FAO) warned last week.

Concern is growing that the Rift Valley in east Africa is particularly at risk. Researchers told *Nature* that if the virus reaches the area's lakes, the health and economic consequences could be even worse than in southeast Asia, where the virus is endemic in many regions.

"The next step we expect the virus to take is into Africa, because that is on the main migratory route for birds," says Ward Hagemeijer of conservation organization Wetlands International in Wageningen, the Netherlands. "The first birds are already in east Africa." There is no definitive evidence that migratory birds are carrying the disease, he says, but the pattern of the virus's spread points strongly to wildfowl travelling southwest from northern Russia to east Africa (see 'Ornithologists on the front line').

Outbreaks have already occurred along the route in Romania and Turkey, where they seem to have been controlled by culls and quarantine orders. But the migration is now in full swing, and some birds have crossed the Middle East and reached the Rift Valley. Middle Eastern states have stepped up efforts to plan for possible outbreaks, and some are monitoring migrating birds, but there has been little response in more vulnerable east African countries. This means that if the virus arrives there, it could quickly become endemic.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Migratory birds could expand the geographical spread of the avian flu virus as far as Africa this year.

"The impact in Africa will be dramatically different from the impact in Europe," says Hagemeijer. Rural communities around the lakes of the Rift Valley region depend heavily on poultry to survive, and live in close contact with both domestic and migratory birds.

The FAO's chief veterinary officer, Joseph Domenech, said on 19 October that African countries should be given international assistance to help them look out for the virus and control any outbreaks. He warned that if the

virus becomes endemic in Africa, the chances of it mutating to spread between humans, potentially triggering a pandemic, will increase.

But the most immediate threat is the economic loss that will result if large numbers of poultry succumb to the disease or have to be culled. "Losing poultry would have a devastating effect on livelihoods in the area," says Lea Borkenhagen, sustainable-living development manager at the charity Oxfam, UK. "Women in particular would be affected, as poultry

Ornithologists on the front line

Geese take off from a pond in rural France at the start of Jacques Perrin's 2001 film *Winged Migration*, which follows the birds until they return the next spring. Perrin tracked dozens of bird species across every continent, capturing the huge scale of the migration phenomenon.

Migratory birds are now being tracked again, but this time because they are suspected of spreading the avian flu virus. The role of wild birds in this has been controversial, and ornithologists are quick to point out that, in many cases, the direction and timing of flu spread is at odds

with what is known about migratory routes. But after recent outbreaks of the H5N1 strain in wild birds in Romania, Turkey and Greece, experts have little doubt that it was carried by birds migrating from Siberia, Kazakhstan and Mongolia, where H5N1 outbreaks occurred earlier in the year.

The worry now is that birds heading from the north to their southern wintering grounds will vastly expand the range of the H5N1 virus — and that their return in spring will bring a new wave of virus to the north. Governments

desperately want to know where and when further outbreaks are likely, so ornithologists have been thrust into the spotlight. Wetland sites across the world have become the new front line against bird flu.

Ornithologists depict the main migration corridors as 'flyways', drawn on maps with breeding grounds at one end, overwintering sites at the other, and the area that birds fly over to get from one to the other (see map). Flyways bundle together the routes of dozens of species, estimated from ringing studies, and give a general picture

of where infected birds might go.

Many birds fly from Siberia to Europe, but the main route is towards southwest Asia and east Africa, says Jan Veen, an expert on Siberian migration routes at Wetlands International. Two other large flyways connect Siberia to the Indian subcontinent, and another goes to the Black Sea and north and east Africa. The recent European outbreaks seem to be along this last flyway, says Veen, suggesting that countries along this route are at risk.

Flyways can also help to show where different species cross paths



Y CHROMOSOMES REVEAL FOUNDING FATHER

Did conquest and concubines spread one man's genes across Asia?
www.nature.com/news

SOURCE: D. A. SCOTT & P. M. ROSE ATLAS OF ANATIDAE POPULATIONS IN AFRICA AND WESTERN EURASIA (WETLANDS INTERNATIONAL, 1996)



Mass migration: red outlines show flyways of the Garganey duck, a species at high risk of spreading the H5N1 virus. The birds breed in the north in summer (dark orange) and then fly south for the winter.

represent the only assets they can possess."

Tadelle Dessie, a poultry researcher at the International Livestock Research Institute in Addis Ababa, Ethiopia, agrees. Earlier this year he travelled to Vietnam where more than 40 people have been killed by H5N1, and spent time with experts working to contain the virus and develop systems to spot it early. "The situation in Africa could be worse," he warns.

There is concern that, so far, not enough is being done to look out for bird flu in the region. "We need to raise awareness of the risks," says Hagemeyer. "Experts know about it, but there is no wider coordination." He is currently travelling in Africa as part of an effort by Wetlands International to establish a network of people monitoring water birds for signs of the disease.

Few countries in the area have systems in place to test for H5N1. And efforts to control outbreaks would be hampered by the difficult terrain in remote communities around the region's lakes, as well as by the low level of education across the region. "Our estimate for the Rift Valley is that literacy rates are between 10% and 40%, with women at the lower end," says Borkenhagen. "That's much lower than in countries such as Vietnam, so it will be hard for people to take on the kind of information that will have to be given."

Dessie says his attempts to get the Ethiopian government to acknowledge the problem have so far been unsuccessful, but that he hopes the FAO's warning will lead to action.

In response to Domenech's statement, Kenya, Sudan and Tanzania have all imposed restrictions on poultry imports, and Tanzanian livestock officials say they are teaching people in wetland areas to keep poultry away from wild birds. East African nations will meet next month in Rwanda to develop a regional strategy to handle the threat of bird flu. ■

and birds in one flyway might infect those on another. Of particular interest are the flyways of ducks, geese and swans, as these come into closer contact with humans than do waders and cranes. They are also frequent carriers of flu, with up to 20% of birds infected.

But for accurate risk assessments, researchers need to look at the routes of individual species and even populations, and to take into account a host of other factors such as bird virology. Until now this has not been a well-funded area, and data are severely lacking. For

instance, vast databases of ringing data covering the past century could provide more precise information, but there has been little money to analyse them, says Jacquie Clarke, head of the ringing unit at the British Trust for Ornithology.

A more high-tech way to track migration uses satellite monitoring of birds fitted with global positioning system (GPS) antennas. On the hub between the Eurasian and North American flyways, birds can be tracked in real time to just a few metres, says Dirk Derksen, chief of the wetlands and terrestrial ecology

branch at the Alaska Science Center.

Until recently, the weight of GPS antennas restricted their use to large birds such as swans, but miniaturization now allows smaller birds to be tagged. But the technique remains very expensive. The Alaska Science Center is also working with the US Air Force on another technique: getting detailed information on the movements of bird flocks from radar data.

But as scientists rush to understand the role of migratory birds in spreading avian flu, those watching birds flying from Europe

to Africa fear that the process is already under way. "Lake Victoria receives huge numbers of migratory birds from eastern Europe and Siberia," says Achilles Byaruhanga, executive officer of Nature Uganda, with some 2 million white-winged terns expected at the Lutembe Bay marshland sanctuary alone.

Byaruhanga is monitoring migration for conservation purposes — this year, he will also be looking out for dead birds. But Uganda, he adds, has no system in place to monitor birds for H5N1.

Declan Butler

ON THE RECORD

“The station is like an old suitcase whose handle is missing... it's totally useless, but you just can't bear to part with it.”

A former space explorer gets sentimental about the ailing International Space Station

SCORECARD



Pine trees
Australian forestry authorities have netted around US\$1 million by auctioning off some ancient trees. Nearly 300 cuttings taken from Wollemi pines, rare fossil trees that grow in a secret grove in the mountains near Sydney, were sold by Sotheby's on 23 October.



Giant squid
Actor and wannabe rock star Kevin Bacon has penned a song to honour the giant squid, photographed last month by Japanese scientists.



Consumption
Disease experts are speculating that tuberculosis and infertility were partly to blame for the gloomy view of humanity displayed by author George Orwell in his dystopic novel 1984.

NUMBER CRUNCH

Hurricane Wilma

On 18 October, Hurricane Wilma screamed into life, just two months after Hurricane Katrina devastated the US Gulf Coast region. By the time Wilma hit the Yucatan Peninsula of Mexico, on 21 October, it had set a record for the strongest hurricane ever to rise from the Atlantic Basin.

882 millibars was the record-setting low pressure reached by Wilma at its peak.

888 millibars was the previous record for Atlantic hurricane intensity, set by Hurricane Gilbert in 1988.

870 millibars was the peak intensity of 1979's Super Typhoon Tip, which holds the record for the strongest and largest cyclone.

Source: US National Oceanic and Atmospheric Administration

P. GIARDINO/CORBIS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Trial aims to measure social effects of choosing babies' sex

WASHINGTON DC

US doctors have launched a clinical trial to assess the effects of allowing couples to select whether they will have a boy or a girl.

Doctors can use a technology called pre-implantation genetic diagnosis (PGD) to examine the sex of embryos that they create by assisted reproduction. Couples then select male or female embryos to implant in the mother's uterus, but the practice is controversial and banned in a number of countries.

Sandra Carson and two colleagues at Baylor College of Medicine in Houston, Texas, started their trial last month, after nine years of consultations with their institutional review board. The doctors have a waiting list of at least 50 couples, but they will only enroll those who already have a child, and want to have a child of the opposite sex — an approach referred to as 'family balancing'.

An experimental technique called sperm

sorting is currently being tested to see whether it can reliably create embryos of a desired sex. But Carson says no one has examined what happens when couples use PGD, a more established tool, to choose an embryo's sex for non-medical reasons.

The practice, known as 'social sex selection', is thought to be common in the United States. One study found that almost 3% of PGD procedures, which are performed thousands

of times a year, were used to choose a child's sex (K. Sermon *et al. Hum. Reprod.* **20**, 19–34; 2005). But countries such as Britain and Canada have banned the practice owing to

public concerns that it could lead to discrimination against women.

"There are still a lot of questions in people's minds about whether this is something that should be pursued," says Robert Brzyski, a fertility doctor at the University of Texas Health Science Center in San Antonio. Brzyski says

"We can sanction and remove members who don't obey ethics statements."

Right to choose? Although private US clinics allow parents to decide the sex of their child, little work has been done on what this means for the family.

he does not offer social sex selection at his clinic because it contradicts the idea that a child should be unconditionally loved, regardless of its sex. "It undermines the principles of the parent-child relationship," Brzyski says.

The United States does not regulate social sex selection, but in the past few years two professional societies, the American Society for Reproductive Medicine (ASRM) and the American College of Obstetricians and Gynecologists (ACOG), have issued statements opposing it.

A spokesman for the ASRM, Sean Tipton, says that the society has no comment on the trial that Carson is running. However, Tipton says, "we can sanction and remove members who don't obey ethics statements."

Carson thinks her study, which will look at the health of the babies born as well as social factors in the families as the children grow up, could convince the ASRM and ACOG to revise their position.

"Their statements are based on public opinions, not outcomes," Carson said at the ASRM meeting in Montreal on 18 October. "Public opinion is important, but it shouldn't be used to ban something."

Erika Check

Europe revamps visa rules to attract world's best minds

The red tape hampering researchers who want to do science in the European Union (EU) has been trimmed. A directive passed this month aims to make it faster for non-EU scientists to get the visas they need.

The move is part of a drive to make the EU more competitive as a knowledge-based economy compared with fast-growing regions such as Asia. "One way to do this was to make it easy for researchers from around the world to work in the EU," says Georges Bingen at the European Commission's mobility directorate.

At the moment, scientists from non-EU countries such as India, China or Iran can struggle for months to secure a visa. The new directive fast-tracks applications from researchers wanting to work for more than three months in an EU country. These 'scientist visas' will also allow researchers working in a Schengen country — one of 15 European states that have abandoned passport control between their borders — to work in other Schengen countries during their stay.

Research organizations will be central to arranging visas by certifying the status of guest researchers in formal hosting agreements, instead of leaving scientists to fight with consulates on their own.

Some research organizations, such as Germany's Max Planck Society, already have departments that help foreign visitors with their visas. "Without help from a hosting institution, non-EU scientists can wait up to a year for a visa, and sometimes arrive with the wrong papers," says Ellena Kempe, who runs the International Office at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden. The new directive should reduce the number of problems, although it won't affect security checks on scientists from politically sensitive countries.

EU ministers have also approved a recommendation to speed up short-term visas for non-EU scientists coming to meetings. "We have heard about Chinese researchers, for example, wanting to attend a meeting in Paris and finding the visa procedures so long and time-consuming that they couldn't attend," says Bingen.

Alison Abbott

Advisers knock Katrina health tests

WASHINGTON DC

The Environmental Protection Agency (EPA) is under fire for the way it responded to the flooding of New Orleans after Hurricane Katrina.

Members of the agency's science advisory board have complained that the EPA had no plan for how to respond to such a disaster and that the first tests it carried out were inappropriate. And scientists and environmental groups are concerned that — nearly two months after the disaster — the agency is still not providing returning residents with information on mid- to long-term health risks.

The EPA, based in Washington DC, was criticized in 2003 for its response to the World Trade Center attack on 11 September 2001. After an investigation into concerns that it had downplayed the respiratory dangers following the collapse of the towers, EPA inspector-general Nikki Tinsley warned the agency that it needed to take responsibility for giving the public reliable advice on health risks after any future disasters. "The EPA needs to be prepared to assert its opinion and judgment on matters that impact human health and the environment," said Tinsley.

But according to the EPA's own science advisory board, the agency has not learned from its mistakes. When New Orleans flooded, the agency did not have an emergency plan in place, and did not test for many immediate health threats, such as the pathogens *Vibrio cholerae*, *Escherichia coli* O157, and *Salmonella* spp., which often contaminate floodwaters. Instead the agency stuck to its normal tests, including measuring pesticides and metals known to cause a long-term risk.

"The EPA would be well advised to put more thought into response plans that are different from the standard operating procedure," Granger Morgan, chairman of the EPA's science advisory board, told *Nature*. "We are going to have a lot more Katrina-like events in the future. It would be good to have a well developed plan for rapid response."

Morgan and other members of the advisory board were so unimpressed with the EPA's performance, they are carrying out their own study into how the agency can develop plans for quick and targeted responses to various kinds of disaster. They will present it to the agency within a couple of weeks.

However the EPA's William Farland defends its response. He says that the agency tested for faecal matter and that even if more specific tests had been carried out, the health advice — "steer clear of the water" — would have been

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The Environmental Protection Agency is accused of failing those returning to sodden homes in New Orleans.

the same. "We would not have done anything particularly different if we had looked at a whole suite of biologicals," he says.

Now that the floodwaters have receded, environmental groups and others are voicing concern about the EPA's ongoing response. The EPA says it is gathering information on short-term risks, looking for levels of chemicals that would be dangerous over time periods of

up to a few days. It has concluded that almost all of its hundreds of samples of water, air and muck would not cause immediate harm. And so far there have been no widespread outbreaks of disease.

But as residents return to the city, many observers feel the agency should be turning its focus to longer-term health risks.

"If you are just looking at a scenario for a

AP PHOTO/K. DIANSEZIAN



POLLUTION MAKES FOR MORE GIRLS

Dirty air skews sex ratios in Sao Paulo, finds Brazilian team.

www.nature.com/news

person who comes in for one day, you come to a different conclusion than if you look at people who are living here," says Gina Solomon of the Natural Resources Defense Council (NRDC), an environmental lobby in Washington DC. Solomon is worried about carcinogens in the soil such as arsenic, which may have leached from building materials, and polycyclic aromatic hydrocarbons, known as PAHs. Such toxins could be present at levels that do not pose a short-term risk, but that far exceed residential standards. "The soil is starting to resemble industrial soil," she says. "If children are coming back to play in those yards, it is really not OK."

Another health concern is the mould rampant in people's houses. "The walls are green and purple and black," says John Pardue, an environmental engineer at Louisiana State University in Baton Rouge. The EPA and other agencies are not testing the mould, saying the priority is simply to get rid of it. But Pardue,

"The soil is starting to resemble industrial soil. If children are coming back to play in those yards, it is really not OK."

Solomon and others are worried that people could be harmed by trying to remove swathes of mould without respirators.

Critics are also upset that the EPA has expressed no opinion on whether it is safe to return to the city, and has said the decision is for local officials to make. "The buck should stop at the desk of the EPA administrator on whether people should return to New Orleans and other affected areas," says Erik Olsen, senior attorney in the NRDC's health programme.

"The EPA has a duty to protect people from known contamination," adds Robert Verchick, professor of environmental law at Loyola College, temporarily headquartered in Houston, Texas, and a research scholar at the Center for Progressive Reform. "It is aware that in some places the contamination exceeds its residential and in some cases two-week standards."

The EPA says it is spreading the word about possible risks through radio announcements, flyers and door-to-door chats. Their website also contains hundreds of air and sediment measurements, although all are presented in the context of short-term exposure.

Earlier this month, the office of the inspector-general at the EPA announced that it will carry out three evaluations of the agency's response to Katrina, one of which will focus on whether good information about safety is being gathered and passed on to the public. ■

Emma Maris

Ministers agree to act on warnings of soaring temperatures in Africa

JOHANNESBURG

African governments need to take urgent action to adapt to global warming. That was the message of 60 researchers who met with policy-makers in Johannesburg last week. Although they did not recommend specific courses of action, the scientists described the ecosystems and people in the continent likely to be worse affected, in the hope that politicians will help work out how to soften the blow.

"We urgently need to determine how we can adapt to climate change, and what the most appropriate interventions should be," says zoologist Steven Chown from the University of Stellenbosch, South Africa.

Temperatures in Africa have risen up to 1°C in the past century and, even if the emission reductions of greenhouse gases agreed by the Kyoto Protocol are achieved, temperatures could rise a further 2–3 °C by 2050, according to climatologist Bruce Hewitson of the University of Cape Town.

Studies into the likely consequences of global warming for Africa are patchy, and so far have tended to focus on South Africa, where science is best funded, and on researchers' existing interests, such as changes to biodiversity. But those at the conference said they are beginning to build a picture of Africa's problems.

One consequence of increasing temperatures is the advance of invasive species. On Marion Island in the southern Atlantic ocean, for example, Chown and his colleagues have shown that indigenous species of springtails (*Collembola*) are declining in number, and invasive ones are increasing. Apparently, the indigenous species prefer cooler temperatures.

Hewitson predicts that reduced rainfall, expected to accompany the rise in temperature, will have severe effects on biodiversity. South Africa's dry west, for example, includes two regions with particularly diverse plant species. These are the Fynbos and the Karoo of South Africa, in which species numbers are predicted to fall as temperatures rise. Wendy Foden, a biologist at the South African National Biodiversity Institute in Pretoria, spoke about the recent decline of the quiver tree (*Aloe dichotoma*). This conspicuous succulent is disappearing in the north of its range in northern

Namibia, and at lower altitudes, as temperatures rise. It cannot establish itself in the cooler south because the rains are not heavy enough.

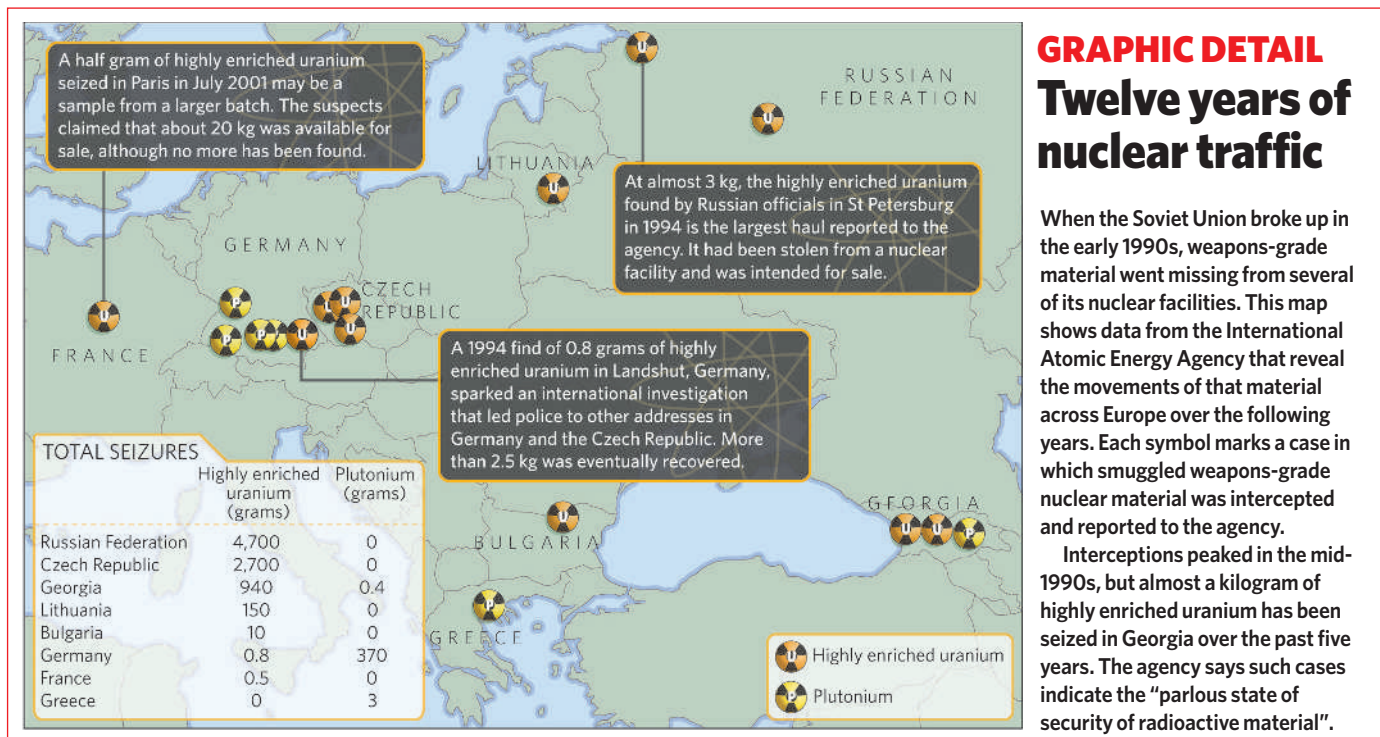
Animals are generally less vulnerable to changing conditions because they can migrate, but those in fenced reserves could be in trouble. Ecologist Norman Owen-Smith from the University of Witwatersrand, Johannesburg, followed roan antelope, sable antelope, eland and tsessebe in South Africa's Kruger National Park. He found that in two decades they had all declined in number and suggests that this may be because inappropriate responses to reduced rainfall have altered the habitat.

As well as being a conservation concern, changes in biodiversity could directly affect fishing, agriculture and tourism. For example, William Bond, a botanist at the University of Cape Town, has documented an increase in tree cover in the eastern part of South Africa over the past 50 years. More trees are bad for tourism as they make game less visible, and bad for farmers because there is less land for grazing. Bond thinks that increased carbon dioxide levels are causing saplings to grow faster, reducing the amount of time that they are vulnerable to bushfires, although this needs to be tested directly.

And others presented evidence that accessible fish stocks have already dropped by 30% in Lake Tanganyika in Tanzania. John Reynolds from Simon Fraser University in Vancouver, Canada, says reduced rainfall means fewer nutrients reach the surface waters.

In South Africa at least, the government seems to be listening. More than 500 policy-makers attended the meeting, including deputy president Phumzile Mlambo-Ngcuka. South African ministers responsible for environment, energy, agriculture and water each promised to outline proposals to mitigate the effects of climate change. "It was wonderful to see the strong commitments made," says Reynolds. "The key is to convert these concerns into effective action."

For African countries with fewer resources and scientists than South Africa, that challenge will be even greater. ■
Michael Cherry



Australia hands cache of fossils back to China

A two-year investigation by Australian police has led to around 10,000 fossils that had been illegally smuggled being returned to China.

The haul is estimated to be worth nearly US\$4 million. It includes undescribed material of scientific importance as well as common specimens such as dinosaur eggs, says John Long, a palaeontologist at the Museum Victoria in Melbourne who assisted the authorities with the fossil inventory. The material was returned late last month. Australian officials are now focusing on possible illegal trade with Argentina, and may launch a similar probe into fossils removed from Brazil, adds Long.

Fossils from China are classed as national treasures that cannot be sold without government permission, but many specimens enter the black market after being unearthed by local people. Although there have been isolated seizures of smuggled fossils in countries around the world, no other nation has undertaken the aggressive approach of Australia.

Depleted stocks prompt call to ban shark fishing

Fisheries researchers have called for a complete overhaul of deep-sea fishing and an immediate ban on catching deep-water sharks.

The International Council for the Exploration of the Sea (ICES), an intergovernmental organization that advises 19 nations, made the call on 17 October after a 20-day survey revealed that shark stocks in some regions are highly depleted. Up to 50 vessels are operating in largely unregulated deep-water fisheries to the west and north of Britain and Ireland, dropping nets that are left unattended for up to ten days. Some nets are ultimately abandoned, and go on to kill large numbers of sharks and other fish.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Sinking fast: stocks of the spurdog shark have fallen to dangerously low levels.

Darwin's body of work evolves into online archive

The complete works of Charles Darwin are to be made available online for the first time. The archive will include 42 volumes written or edited by Darwin and hundreds of shorter publications such as journal articles. Darwin's private notebooks, in which he recorded the observations and thoughts that led to his theory of evolution, will also be digitized. The documents will be available both as searchable text and exact reproductions.

The project, based at the University of Cambridge, UK, will launch its website in December. Project scientists also plan to include translations of some of the works into other languages. The £250,000 (US\$440,000) needed to fund the scheme was provided by Britain's Arts and Humanities Research Council.

http://darwin-online.org.uk

With some species such as the spurdog shark (*Squalus acanthias*) depleted to record lows, ICES scientists report that populations of the more sensitive shark species are in danger of collapse.

Yale hit by discrimination row over Chinese graduate

Graduate students have accused Yale University of discrimination after a graduate in the Department of Ecology and Evolutionary Biology was asked to leave the university because of her poor academic standing.

Xuemei Han maintains that she met all the academic requirements, and claims that members of the department felt uncomfortable working with a Chinese student. The grievance, filed with the dean's office on 20 October, was signed by more than 300 graduate students and scholars. The signatories allege that Chinese graduate students consistently face intimidation, erroneous reporting of academic performance, and other discriminatory treatment.

"This type of unfair treatment of Chinese students happens all the time at Yale: every year, every semester," says Han. University officials deny any wrong-doing and say they will seriously consider Han's complaint.

German researchers angry at plan to expand ministry

Science organizations in Germany have hit out at plans to hand responsibility for some areas of research to new economics minister Edmund Stoiber.

Stoiber — a former prime minister of Bavaria — says he wants to expand his ministry to cover both economics and technology. His plans involve taking

control of research programmes in future technologies such as space and nanotechnology.

The proposal is currently being discussed by the parties involved in negotiations on the make-up of the new coalition government, led by Chancellor Angela Merkel. Scientists fear that splitting research between ministries would make it complicated to run interdisciplinary research programmes and hamper efforts to encourage basic and applied researchers to work together.

"Dividing responsibilities for research between ministries would endanger our ability to be innovative," says Jürgen Mlynek, president of the Helmholtz Association, which runs Germany's 15 national research centres.

Europe's nascent research council takes shape

The proposed European Research Council (ERC) will be run by a scientist, not a bureaucrat, the first meeting of the ERC's Scientific Council has decided.

The decision, made on 18–19 October, is likely to be welcomed by many European researchers. The ERC will be the first Europe-wide basic-research agency, and is designed in part to reduce the bureaucratic burden associated with the European Union's Framework research programme. Some officials involved in the plans for the ERC, which is due to start up in 2007, had wanted it run by a bureaucrat seconded from a national research agency.

Researchers at the meeting also confirmed that the ERC is to fund basic research across all academic fields, and endorsed plans for an executive agency, under the aegis of the European Commission, to implement all ERC activities.



WHILE YOU WERE SLEEPING

The flailing limbs of someone acting out their dreams in bed may not seem the obvious place to seek a cure for Parkinson's disease. But, as **Alison Abbott** finds out, this sleep disorder is shedding fresh light on the development of neurodegenerative disorders.

Many couples suffer in silence, for fear that the police may be called. But those who make it to the sleep clinic have some bizarre tales to tell. Sleep neurologist Brad Boeve recalls one couple who described a particularly horrendous night-time event. While they slept in their bed, the husband suddenly grabbed his wife's head, shook it around roughly, then slammed it down hard and threw up his arms.

Far from being intentional, this distressing episode was the result of a disorder that sees sleeping people physically act out their dreams. In this instance, when the husband woke up he revealed that he had been playing rugby in his dream, had scored a try and then raised his arms in victory.

The case is just one of many that Boeve has investigated at the Mayo Clinic in Rochester, Minnesota. Beyond the obvious immediate trauma, Boeve believes that the sleep problem could hint at something more sinister. A large proportion of those affected by it go on to develop Parkinson's disease or a closely related neurodegenerative condition.

"I get weekly e-mails from people who have had these episodes and are afraid," says Boeve. The idea that the sleeping problem, known as REM sleep behaviour disorder, or RBD, and Parkinson's could be linked poses difficult ethical questions for doctors. But it could also overturn current theories on how Parkinson's disease works and how it could be treated.

There are many sleep disorders, from insomnia to sleep-walking (see page 1253), but few are more bizarre and disturbing than RBD in its chronic form. About a quarter of a healthy night's sleep is occupied by rapid eye movement (REM) sleep, snatched in ever lengthening stretches as the night progresses. REM sleep is the time of dreaming, and voluntary muscles — apart from those of the eyes, which flicker continuously — become temporarily paralysed. Described as REM atonia, this paralysis stops us from physically acting out our dreams.

RBD sufferers lose this atonia. Instead, they flail their limbs and carry out coordinated movements — like the head-grabbing — following the course of their dreams. These

dreams are always vivid and frequently involve fighting or being chased. Experts believe that the subjects of the dreams are, in turn, influenced by the mobilized limbs whose movements get woven into the stories.

Strong link

Episodes of RBD may occur only once a year, or as often as four or five times a night. So far, there seems to be no link between the frequency of the episodes and the likelihood of developing a neurodegenerative disease — or how soon such a disorder might appear. Of the 26 RBD patients studied by neurologists Carlos Schenck and Mark Mahowald at the University of Minnesota in Minneapolis, 18 have developed Parkinson's disease or closely related conditions¹.

The degenerative diseases that have been linked to RBD have a common underlying feature. A protein in the brain called α -synuclein becomes misfolded and clumps together to form aggregates. Collectively, these diseases are called synucleinopathies, although it remains unclear exactly what role the



α -synuclein protein plays in the conditions.

RBD patients develop diseases involving misfolded α -synuclein disturbingly often. Schenck, for example, notes that about 70% of his RBD patients go on to develop a synucleinopathy. In his patients, it took an average of 13 years from the onset of RBD until the patient began to show symptoms of synucleinopathy — although the range was between 3 and 29 years.

Other centres report similar experiences. In unpublished work, Boeve, for example, has studied more than 250 RBD patients, many of whom had Parkinson's disease or another synucleinopathy when they arrived at his sleep clinic. About half of those who showed up with just RBD developed a synucleinopathy an average of eight years later. Tell-tale α -synuclein aggregates called Lewy bodies were found in the brains of 26 of the 27 patients who died during the study.

By the mid-1990s, it was clear that there was probably a connection between RBD and synucleinopathies². But researchers had no idea what might lie behind the link — anatomically it just didn't add up. A muscle's readiness to contract, or its tone, during REM sleep seems to be controlled by the brain stem, a complex structure that shuttles information between the body and higher areas of the brain. This was demonstrated in the 1960s by sleep researcher Michel Jouvet at the University of Lyon, France. He damaged the pons, part of the brain stem, in cats and found that the ani-

mals no longer slept peacefully but would, for example, stalk imaginary prey during REM sleep³.

But according to scientific dogma, Parkinson's disease results from the death of nerve cells in the substantia nigra, a different region that sits above the brain stem in an area called the midbrain. These nerve cells release a chemical called dopamine, which transmits signals to other parts of the brain and so helps to control movement. The link between Parkinson's and the loss of these cells in the substantia nigra is

"I have a feeling that in Parkinson's disease we may have trouble seeing the forest for the tree." — William Langston

reasonably well established. Treating Parkinson's patients with dopamine, for example, can dramatically improve their movement control, at least for a while. And autopsies of people with the disease have shown that they had lost at least 80% of their dopamine-producing cells in the substantia nigra.

So how can this be reconciled with the apparent correlation between RBD and Parkinson's disease? The small community of scientists researching this field are now wondering whether the dogma needs a rethink. They suspect that RBD is the first sign of a degenerative process that begins in or near the brain stem and then creeps up the brain

to other areas. According to this idea, first put forward by Boeve⁴, RBD results from damage to a brain area affected early in the course of a more extensive condition. This disease really cripples only when it hits, and strips, the substantia nigra.

However plausible it seems, this idea leaves many questions unanswered. Why, for example, do only two-thirds of Parkinson's patients seem to suffer from RBD — shouldn't they all? But there is also some compelling support for the concept, particularly from the work of Heiko Braak, a neuroanatomist at the University of Frankfurt in Germany.

Braak performed detailed anatomical investigations of 41 autopsy brains from people who had Parkinson's disease. He also looked at 69 brains from people who had no clinical record of neurodegenerative disease but whose autopsies revealed that parts of their brains contained Lewy bodies⁵.

Body of evidence

Braak showed that the emergence of Lewy bodies seemed to follow a defined and fairly predictable path, which he graded into six distinct steps (see graphic, overleaf). In the least affected brains, the Lewy bodies are confined to distinct areas of the lower brain stem. In stages 3 and 4, damage extends into the upper brain stem. And in stages 5 and 6, the damage reaches the substantia nigra and, ultimately, the cortex, impinging on areas involved in emotional and intellectual activities⁵. The Lewy bodies seem to multiply and spread through the brain, never appearing at higher levels unless they are also present lower in the brain stem. "By the time you see the distressing motor symptoms of Parkinson's, the damage is well advanced," says Braak.

But because the early stages of this condition don't seem to have any obvious symptoms, it is very difficult for clinicians to correlate Braak's stages with signs of disease. So it is impossible categorically to conclude that the progression of the Lewy bodies has anything to do with Parkinson's. "It is only a hypothesis, although a plausible one" Braak says. "We'll have to see if any symptoms are eventually associated with early stages."

His idea certainly resonates with clinician-researcher William Langston, who heads the Parkinson's Institute in Sunnyvale, California. Langston is keenly aware that his patients report all manner of weird symptoms that at the moment are not classified as part of classical Parkinson's. "There is a ringing association on the clinical side," he says. "I have a feeling that in Parkinson's disease we may have trouble seeing the forest for the tree — the tree being the devastating motor effects."

What would it mean if the early stages identified by Braak were indeed relevant to the disease? And what if the march of the Lewy bodies could be tracked in the living brain, for example through unexpected clinical manifestations such as RBD? "Then we could

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

The loss of motor control in patients with Parkinson's disease may be the late-stage symptoms of a long-term degenerative condition.

give people advance warning that they were at very high risk for Parkinson's," says Ilonka Eisensehr, a neurologist now in private practice whose research at the University of Munich has added to the body of evidence linking RBD with synucleinopathies.

Eisensehr found that some patients who came to her sleep clinic for reasons unrelated to RBD had some muscle tone during REM sleep. In other words, they no longer had complete sleep paralysis, although they had not yet begun acting out their dreams. She carried out brain imaging studies on a group of these 'sub-clinical' RBD patients, comparing them with RBD patients, those with Parkinson's disease and healthy controls. She measured levels of proteins called dopamine transporters in the upper brain stem. These proteins are found only on dopamine-producing neurons, giving Eisensehr a way of monitoring the fate of these cells in patients.

She saw a clear pattern: the greatest loss of dopamine transporters was in patients with Parkinson's disease, although RBD patients also showed fairly low levels. In the subclinical RBD patients, the transporter level was better but was still lower than in the healthy controls, and the loss correlated directly with how much muscle tone they had during REM sleep⁶. "Loss of REM atonia could indeed be a measurable, very early indicator for Parkinson's disease," Eisensehr says.

Karin Stiasny-Kolster at the University of Marburg in Germany has taken this further by showing that nearly all of her 30 patients with clinical or subclinical RBD had a badly disturbed sense of smell, a very common complaint among Parkinson's patients⁷. These observations fit well with Braak's hypothesis that Parkinson's progresses up through the brain in distinct stages, says Stiasny-Kolster. Smell signals enter the brain directly in the olfactory bulb, one of the areas, along with the brain stem, that Braak categorizes as stage 1.

Despite this evidence, some sleep neurologists have yet to be convinced. One of them is

Alex Iranzo at the University of Barcelona in Spain. He says that over the past nine years, 40% of his RBD patients have developed a neurological disease, but in many cases that disorder was not a synucleinopathy. "RBD is certainly an anatomical disease, but I'm not sure there is a simple molecular explanation for it," Iranzo says.

Clearly the phenomenon requires a lot more research to unravel exactly what is going wrong in the brains of those who suffer from RBD. But whatever the outcome, the statistics are stark: these patients have a highly increased risk of developing Parkinson's disease, or another irreversible degenerative disorder. At the moment there are no drugs to protect neurons against the disease, raising the ethical dilemma of whether or not patients should be warned when they are diagnosed with RBD.

Policy of truth

Like Eisensehr, Schenck thinks they should be. "Then they will stay in touch with clinics and may eventually be able to be recruited into studies testing potential neuroprotective agents," Schenck says. Many drug companies are trying to develop such drugs, and enrolling people at high risk of developing Parkinson's, but who have no symptoms, into clinical trials could help bring products to the market faster. Recruiting those who already have Parkinson's disease is not ideal as they have already lost a high proportion of their dopamine-producing cells, leaving few neurons to protect. "Most trials have failed because we are not getting in early enough," says Langston.

Jacques Montplaisir, a neurologist at the University of Montreal in Canada, prefers not to bring the issue up with his patients unless one of them specifically asks. He is concerned about needlessly alarming the subgroup of RBD patients who are not suffering from a neurodegenerative disease. Montplaisir tests his RBD patients' sense of smell as well as their ability to discriminate colours, which is also

frequently lost in Parkinson's disease, hoping that this will help to identify those who are at high risk. Only at such a point, or when a neuroprotective agent is available, would it be ethically justified to inform patients, he says.

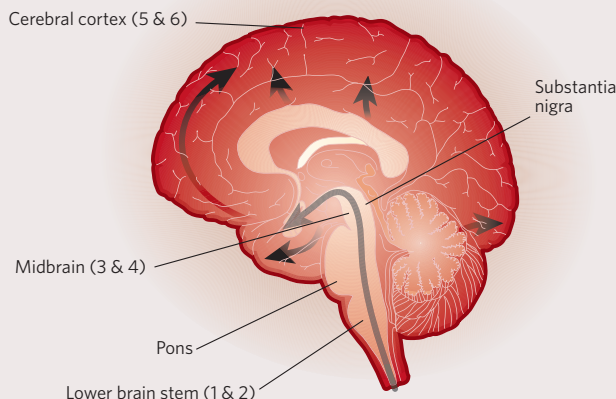
"The best approach is to make clinicians aware of the possible correlation between RBD and Parkinson's, so they will be alert to danger signals such as RBD, and also disturbances to the sense of smell," says Kieran Breen, director of research at the Parkinson's Disease Society in London. "Clinicians could then monitor the patients closely." But Breen adds that patients should be told they have Parkinson's only when the first motor symptoms start. "It would not be fair on the patient otherwise," he says, "given the uncertainty of the link and the long time-lag that can sometimes occur between onset of RBD and onset of Parkinson's."

Neurologists currently prescribe clonazepam, a drug used to treat epilepsy, to RBD patients to suppress their dream enactment. But no prescription can help patients live with the fear that a diagnosis of RBD brings. In the future, those with early warning signs may turn out to be the lucky ones if effective neuroprotective agents are developed.

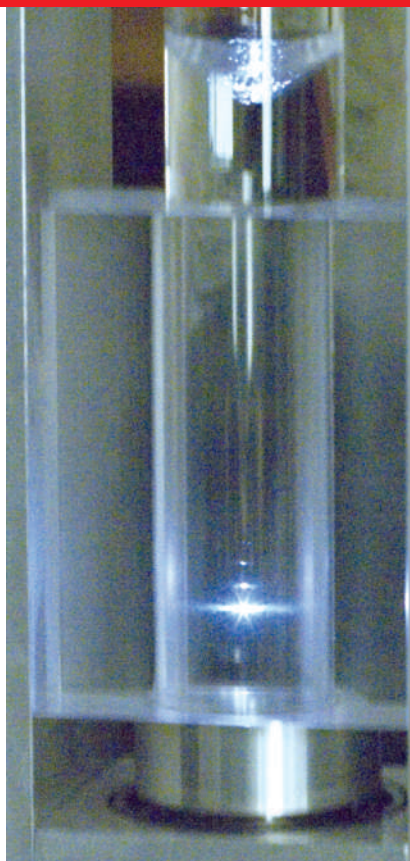
"We are facing a revolution in Parkinson's disease," says Langston. "We've been fixated on movement disorder linked to cell death in the substantia nigra, but we may now have to think more broadly." He believes that early warning signs such as RBD may turn out to be exactly what is needed to understand this terrible disease and learn how to stop it in its tracks. ■

Alison Abbott is Nature's senior European correspondent.

A FRESH VIEW OF PARKINSON'S DISEASE



Parkinson's disease results from the loss of neurons in part of the brain called the substantia nigra. Researchers now suggest that its symptoms are a late sign of a more extensive disease that begins in the brain stem and spreads throughout the brain in six stages.



FAR FROM THE FRONTIER

Ignoring the mainstream, physicist Seth Putterman has a knack for bringing long-forgotten mysteries back to the fore. **Geoff Brumfiel** discovers some of the payoffs, and perils, of being a fiercely independent researcher.

Lawrence Crum thought the experiment was cute; Seth Putterman thought it was miraculous. The year was 1989, and Putterman, a theoretical physicist at the University of California, Los Angeles, (UCLA) was visiting Crum's laboratory at the University of Mississippi in Oxford. One of Crum's graduate students, Felipe Gaitan, was exploring an unusual phenomenon: the creation of light from sound. He began with a glass cylinder filled with a water and glycerine mixture. By vibrating the cylinder at a low frequency, he could create a single bubble that would rhythmically expand and collapse, releasing a tiny flash of light as it did so.

Putterman took one look at a video recording of the flashing bubble suspended in the tube and became feverish. "You shouldn't be able to do it," he remembers thinking. "The energy of a piece of sound is 12 orders of magnitude smaller than the energy of a piece of light." Putterman begged Crum to carry out more experiments on the little bubble. How

did it form? Why was it so stable? Where did the energy for those light flashes come from?

But Crum had other priorities, not least securing adequate funding for the 150 people then working on physical acoustics in Mississippi. The technical term for what Crum's graduate student had seen was sonoluminescence, and it was nothing new — physicists had known about it for at least 60 years. In 1989 sonoluminescence was not a particularly hot topic. Crum shrugged off Putterman's interest, saying if he wanted to study the little flashing bubble, he was welcome to go ahead and try.

Theoretical physicists spend most of their time at their desks, so perhaps Crum didn't expect Putterman to take up the offer. But back in California the theorist quickly set up his own experiment to recreate the Mississippi effect. He then followed a recipe that has often served him well since: he used precise instrumentation to characterize the phenomenon in detail, and applied a theoretical understanding that was strong enough to challenge popular models of the time.

Within two years Putterman had some surprising results¹. The inside of the bubble was imploding faster than anyone had thought, and the temperature at its core seemed shockingly high — perhaps five times that of the surface of the Sun. Overnight, sonoluminescence

became a cottage industry in the world of acoustical physics. Everyone wanted to know what made it tick.

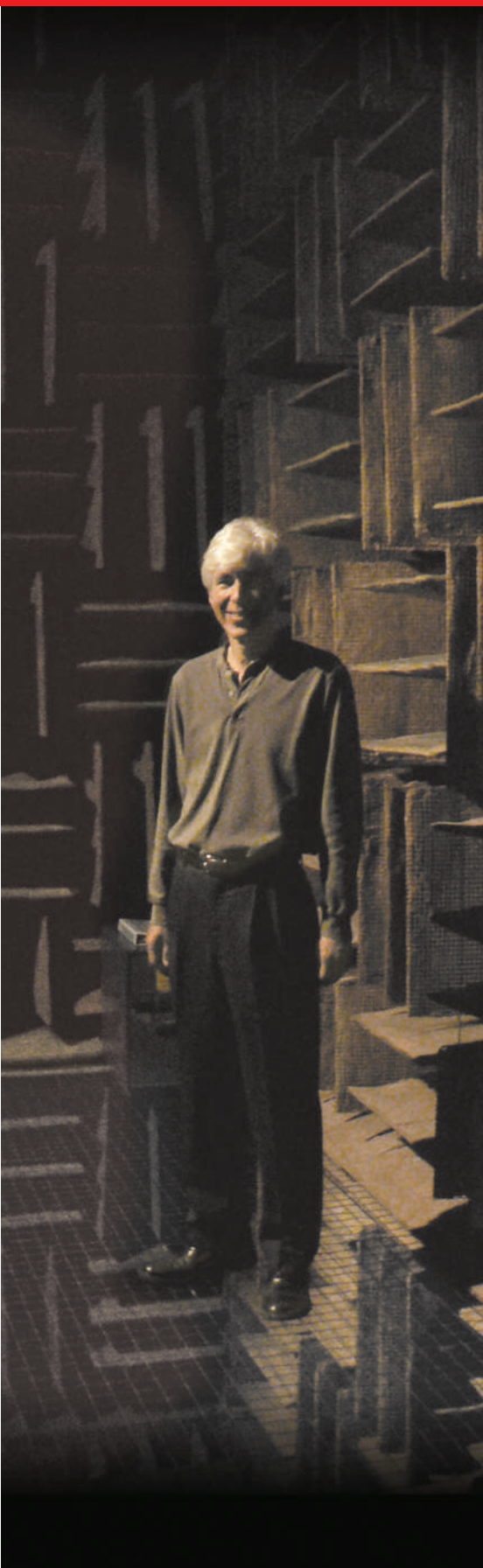
Now 59 years old, Putterman is tall and skinny with a deep Californian tan and a shock of bright, silver hair. Among friends and foes alike, he is known for his love of physics, his penchant for controversy, and his fierce independence. In an era when much of physics advances through the cooperative efforts of hundreds of researchers, Putterman is a lone figure willing to study any phenomenon, no matter how obscure, and ready to publicly take on his critics, no matter how impolitic.

Bright spark

Aside from physics, Putterman loves good wine, and over a bottle of 2001 Killibinbin Shiraz, he recalls his beginnings in science. From an early age, he found himself gravitating towards physics; he studied the subject first at Cooper Union in New York, and later at the California Institute of Technology (Caltech) in Pasadena, where he completed his undergraduate degree in 1966.

Even in those early days, Putterman was unafraid to snub convention. His first grant proposal was drafted to support his love for wine-tasting, not physics. Under the 1958 National Defense Education Act, which encouraged US students to pursue careers in science, Putterman was eligible for a low-interest loan. The government was persuaded by Putterman's argument that although he was sure he wanted to embark on a career in physics, he felt a hobby would help keep him going. "Caltech was very supportive of my

"Seth Putterman is sure of himself, enthusiastic, smart, thinks fast on his feet — and he can be abrasive as hell." — Kenneth Suslick



Lone explorer: Seth Putterman is happy to study phenomena that others dismiss as uninteresting.

proposal,” Putterman recalls. “I got US\$2,000, which went a long way to buying great Bordeaux like a 1959 Lafite and 1959 Yquem!”

After graduating from Caltech, Putterman returned to New York City to attend Rockefeller University, where he studied under George Uhlenbeck, who together with Samuel Goudsmit, discovered electron spin in 1925. Uhlenbeck was a kindred spirit, a man who loved physics and had little patience for the politics of the field. “He had no concern about what was or wasn’t the official frontier of physics,” Putterman reflects.

Burst bubble

But his adviser could also be intimidating. On Saturday mornings, Putterman would often find Uhlenbeck in his office, smoking a cigar and meditating in a large leather-backed chair. “I would tell him some scientific insight I’d had in the last week, and he would go over it, see what it meant, formulate it, work with it, attack it — but he’d be enjoying it the whole time,” Putterman says.

Putterman cherishes his cloistered days at Rockefeller University, but admits that they left him unprepared for the realities of modern science. “I was totally sheltered at the Rockefeller,” he says. The faculty didn’t have to get funding, and Uhlenbeck never discussed university politics with him.

But at least Rockefeller gave him the strength and independence to pursue his own interests, he says. He never worries about ‘frontier science’, as he calls it, even asking his students to replicate a 300-year-old experiment whose results were never explained. In the case of sonoluminescence, this strategy propelled Putterman to celebrity status. But it has also led his critics to accuse him of profiting from recycled ideas. “On a grant proposal I once wrote, I was accused of being a used-car salesman,” he says with a chuckle. “But I think there’s just wonderful stuff in the old literature waiting to be exploited with modern instrumentation.”

Putterman also likes a good fight, and he is never afraid to publicly take on his critics. “He is sure of himself, enthusiastic, smart, thinks fast on his feet — and he can be abrasive as hell,” says Kenneth Suslick, a chemist at the University of Illinois at Urbana–Champaign. Putterman has clashed repeatedly with other scientists over interpretations of his first sonoluminescence experiments. Putterman firmly believes that the flash at the centre of the bubble is created by electrons being shaken out of their atomic orbits, whereas his opponents suspect more conventional chemistry is the culprit. The debate has turned so acrimonious that some of his opponents refused to speak to *Nature* for this article.

Although Putterman is upset by some of the personal disputes that his work has sparked, overall he remains unfazed by funding difficulties and academic controversy. And true to form, earlier this year, his latest

experiments² landed him in the centre of another political tussle.

In his lab at UCLA, Putterman shows off his latest marvel, a crystal just a few centimetres long; when the crystal is heated, an enormous electric charge builds up on its surface. Like so many of the phenomena studied by Putterman, this type of crystal is not new — it was first described by the Greek philosopher Theophrastus in 314 BC. But Putterman is the first to exploit its properties for nuclear fusion.

Putterman is using the crystal’s electric field to catapult hydrogen ions onto a hydrogen-filled target. The result is the fusion of nuclei, which produces helium, and a flurry of neutrons. What is surprising is that this tabletop experiment generates the sort of fusion event that usually requires heavy-duty particle accelerators.

Desktop fusion

The problem with reports of tabletop fusion is that for most scientists they evoke memories of the notorious, and now largely discredited, ‘cold fusion’ claim made by two chemists in 1989. The chemists claimed they could achieve nuclear fusion reactions well below the extreme temperatures predicted by theorists, and that these reactions could be used as a source of unlimited energy.

Perhaps unsurprisingly, the press were quick to label Putterman’s recent findings using the crystal² as ‘cold fusion’. Many scientists might have been horrified to have their research mischaracterized in this way, but Putterman was unperturbed by the controversy, and even enjoyed it. “If people think this is a crackpot paper that’s just fine,” he told *Nature* at the time of his group’s announcement. “We’re right.” In the end, researchers found Putterman’s measurements of the fusion reactions convincing, and Putterman is now planning to develop commercial and medical applications of his work.

Despite turning 60 this year, it seems unlikely that Putterman will narrow his eclectic pursuit of physics. These days, some of his students are attempting to back up the unconfirmed claim that sonoluminescent bubbles are powerful enough to be used as a form of tabletop ‘bubble fusion’³. Another team is starting to make precise measurements related to the randomness inherent in quantum mechanics. It’s an area that has long tickled his fancy, although he knows little about it.

During his long career, Putterman has picked up no society awards, and funding has often been tight, but he continues to be driven by the challenge of science itself. “I don’t do physics with the goal of scoring grant money or proving myself; I do it for the fun of learning something new,” he reflects. ■

Geoff Brumfiel is *Nature’s* physical science correspondent based in Washington DC

1. Barber, B. P. & Putterman, S. J. *Nature* **352**, 318–320 (1991).

2. Naranjo, B. et al. *Nature* **434**, 1115–1117 (2005).

3. Taleyarkhan, R. P. et al. *Science* **295**, 1868–1873 (2002).

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Pipe dreams

It's a bizarre, toxin-filled microbe that could clean up sewage plants across the globe. **Helen Pilcher** gets on the trail of the anammox bacterium.

When the good citizens of Delft complained about the smell wafting from the local yeast factory, they had no idea they had initiated the discovery of rocket-fuel-filled bacteria that would mystify microbiologists and force a rewrite of the biology textbooks.

The microbes, known as anammox bacteria, are full of surprises. They flout bacterial convention by having tiny, sac-like compartments. They defy common sense by packing these sacs with toxic hydrazine. And they are helping to resolve a long-standing conundrum about where much of the ocean's ammonia goes. Not content with this, the brightly coloured bacteria are paving the way for environmentally friendly sewage treatments.

The Gist-Brocades yeast factory, situated in the middle of Delft in the Netherlands, was unpopular with the locals because it produced a lot of eggy-smelling, sulphide-rich waste. To keep its neighbours sweet, the company devised an odourless process, breaking down the effluent in sealed, oxygen-free tanks. The pilot plant built in the mid-1980s worked, and sulphide levels fell. But as the residents breathed a sigh of relief, workers at the plant noticed something odd. Dogma had it that ammonia needed oxygen to be broken down so the engineers expected the concentration of that compound in the tank to stay constant. But a few months later, ammonia levels were falling and nitrogen gas was being produced.

Intrigued, the company contacted microbiologist Gijs Kuenen at Delft University of Technology. Kuenen suspected that anaerobic bacteria were at work, combining ammonia with nitrite to form nitrogen gas and water.

The idea that bacteria could use an anaerobic ammonium oxidation or anammox reaction had been proposed some ten years earlier¹, but most microbiologists were sceptical — no such bacteria had ever been found and the reaction had never been seen occurring naturally.

Kuenen realized that the mysterious Delft bacteria might offer a new method for wastewater treatment and, if found elsewhere, could be important in nature. He decided to investigate². "It was a brave move," says his former PhD student Marc Strous, now at Radboud University in Nijmegen, the Netherlands. "Kuenen began to study something that all his colleagues thought didn't exist."

Vigorous hybrid

Electron microscopy helped to nail the suspect. A close look revealed microbes that housed a strange, internal, membrane-bound compartment. This was a big surprise, because only more complex, or 'eukaryotic', cells such as our own are supposed to have such compartments, called organelles. Simpler 'prokaryotic' cells, including bacteria, should lack them. Only one kind of bacterium, the planctomycetes, was then known to play host to such a structure, and it turned out the microbes belonged to this phylum.

Planctomycetes are peculiar because they contain features from all three domains of life — bacteria, eukaryotes and archaea — and some think they may represent an early com-

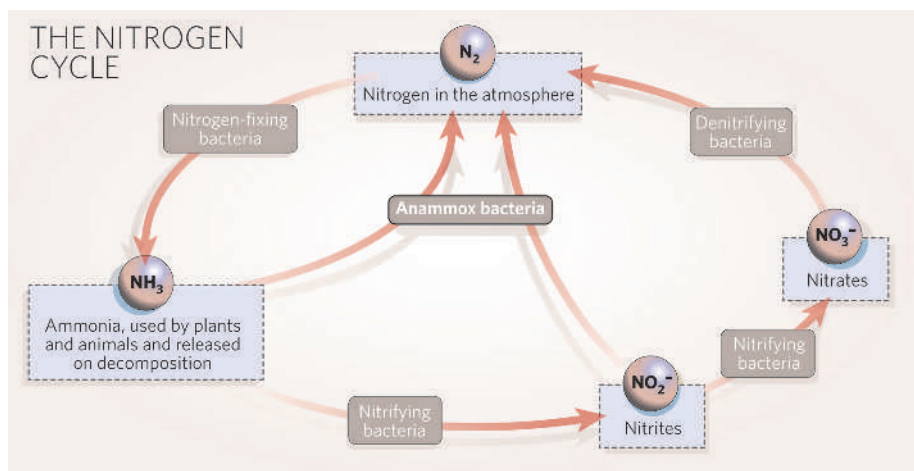
mon ancestor. DNA studies place them firmly in the bacterial camp. But their internal organelles make them resemble eukaryotes. And the microbes lack the rigid polymer peptidoglycan in their cell walls, making them similar to members of the single-celled archaea domain. "They blur the definition of what bacteria are," says Strous.

Planctomycetes were not known to perform the anammox reaction, but Kuenen's team incubated the Delft cells with ammonium and nitrite, and watched the ingredients disappear³. Genetic analysis confirmed the microbes' group and they were provisionally named *Brocadia anammoxidans*; *anammoxidans* for their unique biochemistry, and *Brocadia* for their place of discovery and because the bright red bacteria reminded the researchers of gaudy brocade fabric.

In the wake of this paper, colleagues' opinions changed overnight. "This was a real turning point," says microbiologist Mike Jetten, also from Radboud University, who contributed to the work. Before the paper, the majority of microbiologists did not believe that anammox occurred. After it, the theory gained widespread acceptance and anammox bacteria took their rightful place in Earth's nitrogen cycle.

The nitrogen cycle converts stable nitrogen gas into more usable forms, such as ammonia and nitrate ions, and back again, maintaining a global balance (see graphic, overleaf). Nitrogen gas is converted directly to ammonia by nitrogen-fixing microbes, such as those associated with certain plant roots in the soil. Plants and animals consume ammonia, which is released when they die and decompose. The

"We showed it to other chemists and they said it was impossible."
— Jaap Sinninghe Damsté



next step is the nitrifying bacteria and archaea, which transform ammonia to nitrites and nitrates, and the cycle is completed as denitrifying microorganisms convert nitrates into nitrogen gas, replenishing the atmosphere. Anammox takes a short cut through the cycle, creating a path from ammonia and nitrite directly to nitrogen gas.

The fact that these bacteria pull off such a stunt is remarkable enough. But when researchers investigated how they did it, there were even more surprises. Findings suggested that the anammox reaction occurring inside the membrane-bound bag, or anammoxosome, produced hydrazine as an intermediate⁴. Why would bacteria produce hydrazine — a potent rocket fuel? The explosive molecule isn't found anywhere else in nature. "We were still puzzled as to what was going on," says Jetten.

Toxic cargo

It may be that the high-energy hydrazine is needed to drive the anammox reaction. But this does not explain how the bacteria manage their toxic load without killing themselves. Hydrazine can easily diffuse through cell membranes, so Jetten suspected that the anammoxosome membrane must be unusual and in some way able to contain the dangerous load.

He contacted lipid expert Jaap Sinninghe Damsté from the Royal Netherlands Institute for Sea Research in Texel and they analysed the organelle's membrane. What they found was extraordinary⁵. "We showed it to organic chemists at the University of Amsterdam and they said it was impossible," says Damsté.

The membrane's lipids are made of five carbon-based rings fused together to form a dense ladder. This 'ladderane' lipid is peculiar because it has a lot of energy built into it and is very unstable. It is thought the structure makes the membrane exceptionally dense, and so stops hydrazine leaking into the rest of the cell. "It's a complete mystery how nature makes this lipid," says organic chemist and Nobel laureate Elias Corey from Harvard University, who has made the structure in the laboratory. Researchers are now unravelling the bac-

terium's genome to work out how it does it. The Dutch team, which has patented the production process for the lipid, hopes that the microelectronics industry will find a use for the impenetrable membrane.

The most pragmatic application of anammox bacteria lies in wastewater treatment. Sewage plants and industrial processes, such as fertilizer manufacture and petroleum refining, generate millions of litres of ammonia-rich waste, all of which needs to be broken down. Conventional methods use nitrifying bacteria to convert ammonia into nitrite and nitrate, and then denitrifying bacteria to yield nitrogen gas. The microbes need oxygen, so huge, electricity-gobbling machines are needed to aerate the sludge. And the denitrifying bacteria need an energy source, such as methanol, which they burn to produce carbon dioxide. The process is costly, takes up a lot of space and is unkind to the environment.

Anammox wastewater plants offer major advantages. Anammox bacteria use ammonia as their fuel — there's no need for expensive methanol. They do not need oxygen, so the process uses less electricity. And instead of producing carbon dioxide, anammox bacteria consume it, so the method is environmentally



friendly. Altogether, this leads to a 90% reduction in operational costs and a 50% reduction in space, compared with conventional methods.

The Dutch company Paques, based in Balk, has developed the first anammox reactor. The prototype has been set up as part of a municipal wastewater treatment plant in Rotterdam and is performing well.

It is likely that anammox will become an important part of wastewater treatment, but its role in the wider world is potentially far greater. Oceanographers looking into anammox reasoned that if the reaction went on in anoxic water tanks, it might also occur in oxygen-poor parts of the sea where it could contribute to the oceanic nitrogen cycle. If that were the case, it would solve a 40-year-old marine mystery.

In the mid-1960s, Francis Richards from the University of Washington in Seattle noted that ammonia was inexplicably missing in an anoxic fjord⁶. He proposed that it must be being oxidized anaerobically to nitrogen gas, either inorganically or by some unknown microbe. At the time, oceanographers found the idea preposterous.

But in December 2001, biogeochemist Marcel Kuypers from the Max Planck Institute for Marine Microbiology in Bremen, Germany, and his colleagues went fishing for anammox bacteria in the Black Sea — the biggest anoxic basin in the world.

The team drew samples from between 85 and 100 metres down, where oxygen is absent and ammonia is found only in trace amounts. As suspected, anammox bacteria were present — the first time they had ever been found in the sea⁷. The microbes are incredibly efficient, and it is thought that anammox bacteria may account for up to half of all the nitrogen production in the world's oceans^{8,9}. This is forcing a major rethink of the global nitrogen cycle and slowly persuading oceanographers that denitrifying bacteria are not the only microbes responsible for nitrogen production.

As acceptance of anammox bacteria grows, so too does their grip on the planet. The microbes are turning up everywhere — in fresh and salt water, open oceans and marine sediments, and in wastewater treatment plants all over the world. "One day you discover a bug," says Kuenen, "then ten years later they turn out to be everywhere and important on a global scale. They may even be hiding in the sewer system under your kitchen sink."

Helen Pilcher is a freelance science writer based in London.

1. Broda, E. *Z. Allg. Mikrobiol.* **17**, 491–493 (1977).
2. Van de Graaf, A. A. *et al. App. Env. Microbiol.* **61**, 1246–1251 (1995).
3. Strous, M. *et al. Nature* **400**, 446–449 (1999).
4. Van de Graaf, A. A., de Bruijn, P., Robertson, L. A., Jetten, M. S. M. & Kuenen, J. G. *Microbiology* **143**, 2415–2421 (1997).
5. Sinninghe Damsté, J. S. *et al. Nature* **419**, 708–712 (2002).
6. Richards, F. A. in *Chemical Oceanography* (eds Riley, J. P. & Skirrow, G.) 611–645 (Academic, London, 1965).
7. Kuypers, M. M. M. *et al. Nature* **422**, 608–611 (2003).
8. Dalsgaard, T. *et al. Nature* **422**, 606–608 (2003).
9. Devol, A. H. *Nature* **422**, 575–576 (2003).

BUSINESS

Patent reform prompts intellectual tug-of-war

Reform of the patent process in the United States is shaping up as a battle of wills between the software and biotechnology industries. The outcome has global consequences, as **Emma Marris** reports.

Two draft bills for patent reform are circulating in Congress, with the chance that one of them could pass into law as soon as next month. And the whole world of patents could be shaped by the reforms.

"The United States has been the experimental test bed for reform," says Francis Gurry, deputy director for patents at the World Intellectual Property Organization in Geneva, Switzerland. "The issues there often turn out later to be issues for the rest of the world. So the world watches them closely."

The reform plan has brought two of the industrial sectors that rely most heavily on intellectual property into conflict.

Biotechnology companies make their money by licensing out their inventions, or by using promising patents to attract venture capital. They spend a lot of money defending these patents — but have no strong incentive to change the current system.

The computer-software industry, on the other hand, is growing restless about 'patent trolls', who buy up software patents with the aim of filing lawsuits and extracting licensing fees or settlements from companies. The Washington DC-based Business Software Alliance has therefore been leading the charge for reform — and pushing for changes that would limit the rights of those suing for patent infringement.

The two draft bills under consideration draw heavily from recommendations made by the National Academy of Sciences, the Federal Trade Commission and the US Patent and Trademark Office (PTO) itself.

They contain several provisions that have broad support. One such change is a 'first-to-file' provision, which would grant patent rights to the first inventor to file, rather than to whoever can prove they had the idea first. Although it might seem fairer that the person who thought of the idea first should get patent rights, determining who this is can require lengthy and costly adjudication. A change

would bring the United States into line with the world's other main patent offices, in Europe and Japan — and remove a major obstacle to the global harmonization of patent systems.

Universities and independent inventors have expressed concern in the past that first-to-file would favour large corporations that can file patents fast, and in volume. But Gerald Mossinghoff, a former commissioner of the PTO who now works as a patent lawyer near Washington DC, says that when he totted up the winners and losers of such adjudications, he found that the large parties gained no advantage.

Both versions of the reform bill would retain another quirk of the US patent system — the one-year grace period. Under other systems, inventors are required to file for their patent at the time that they first disclose their invention to the world. In the United States and a few other countries, data on an invention can be published or presented up to a year before a patent is filed.

Joseph Straus, director of the Max Planck Institute for Intellectual Property, Competition and Tax Law in Munich, Germany, says that he hopes Europe will intro-

duce the grace period in exchange for the United States accepting first-to-file patenting. He says the grace period helps innovation by letting scientists speak and write more freely about their work.

The draft bills also install a new vetting procedure called 'post-grant opposition', which would allow a patent to be contested in the first nine months after its approval, without a court hearing. The proceeding would take place in the patent office and the burden of proof would rest with the contestor. Under present rules, companies dubious about the legitimacy of a competitor's patent must wait until the patent holder complains that they are infringing it. Only then can they make a counter-complaint that the patent is no good. Sometimes firms just avoid researching in the area of the patent.

Most agree that post-grant opposition could

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Fruitful enterprise: patent law drives innovations from the Red Delicious apple to the microchip.

be a useful way of cutting down on litigation and improving patent quality. Earlier drafts of the bill allowed a second window for such challenges, in the six months after the patent holder sues someone for infringement. But biotechnology companies, in particular, feared the idea of patents with an air of reversibility about them, and successfully opposed its inclusion.

Capital attraction

Robert Chess, head of Nektar Therapeutics, a California-based biotechnology firm, says that early post-grant review could help innovative companies to attract capital. "We need certainty and early certainty, so investors will be willing to invest," he says. Under another provision, competitors' input could also be submitted during the patent examination process.

One early version of the bill, favoured by the software industry, included a provision that markedly curtailed a patent holder's right to stop another company that was infringing their patent from selling the invention. The industry hoped this would help them deal with their litigation woes. But such a provision would have been anathema to biotechnology firms, and it was eventually dropped.

The main outstanding disagreements concern whether damages should be calculated from the value of the whole product, or just the portion of it covered by the infringed patent, and rules that would limit the courts in which such infringement trials could be pursued. In both

J. LEVINE/CORBIS

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

cases, software companies want to make things harder for those claiming infringement, whereas the biotechnology and pharmaceutical industries want to make it easier.

Despite these areas of dispute, Mossinghoff says he thinks a bill could be passed next month, before Congress packs up for the year. Nicholas Godici, another former PTO commissioner who also now works as a patent lawyer near Washington DC, says there is "a 50–50 chance" that the bill will pass either this year or next.

But the bill won't set funding for the over-worked PTO. As the number of patent applications increases, the office is stretched to its limit, and some observers say that the quality of patent decisions is suffering as a result. The PTO calls the situation a "workload crisis" in its strategic plan, but denies that its standards are dropping. "Questions of bad quality are overblown and anecdotal," claims Brigid Quinn, a spokeswoman for the PTO.

If there was single filing and examination of patents for the United States, Europe and Japan, the strain in each region would be considerably reduced. "The three major offices in the world are beginning to buckle under the workload, so we really have to get cooperation," says Mossinghoff. "This deep harmonization will remain theoretical until we get first-to-file in the United States."

Failure of the US bills would spell trouble on that front, declares Eugen Stohr, head of international affairs at the European Patent Office in Munich, "because then the rest of the world will know that the United States is not able to move to first-to-file — and that discussions on deep harmonization are useless". ■

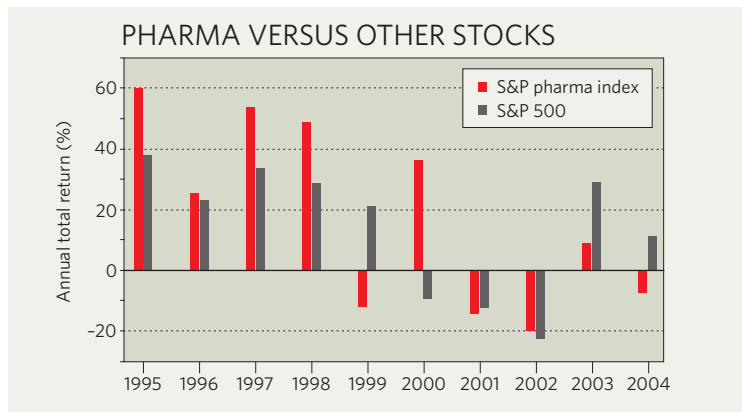
IN BRIEF

WAR ON WORDS The United States' biggest book publishers are taking the Google Internet search engine to court, charging that the web wizard's plan to scan their products and put them online (see *Nature* 433, 446; 2005) would breach copyright rules. The Association of American Publishers filed suit in the US District Court in New York on behalf of five major publishers, including McGraw-Hill and John Wiley, after talks with Google broke down. In a statement, the California-based company called the suit "short-sighted".

TB DRIVE A foundation backed by Microsoft boss Bill Gates has announced a partnership with drug manufacturer GlaxoSmithKline to expedite trials of a tuberculosis vaccine. GSK Biologicals of Rixensart, Belgium, is already conducting clinical trials of a recombinant protein vaccine that seems to induce strong, long-lasting immunity to the disease. The Aeras Global TB Vaccine Foundation will provide up to \$13 million over two years to help speed up the trials. Last year the foundation, which aims to develop an effective vaccine for tuberculosis within ten years, received \$83 million from the Seattle-based Bill and Melinda Gates Foundation to support its work.

FRAUD SQUAD A subsidiary of the Swiss biotechnology company Serono is to pay the US government \$704 million to resolve criminal charges and civil allegations related to fraudulent marketing of Serostim, a drug that treats the wasting induced by AIDS. Serono has pleaded guilty and last week agreed to pay a \$137-million criminal fine and \$567 million to settle civil allegations that it defrauded the Medicaid programme for poor people and the Medicare health insurance scheme by using illegal means to boost sales of Serostim between 1996 and 2004. The settlement is the largest ever in a Medicaid fraud case.

MARKET WATCH



Pharmaceutical stocks have seriously underperformed on the general US stock market for the past two years, as the sector loses some of its cachet. In 2003, for example, when the broadly based Standard & Poor's 500 index (S&P 500) produced an annual return of 29%, pharma managed less than a third of that. Last year, pharma actually lost ground while general stocks showed a return of 11%.

Why the fall? Experts chalk it up to a dearth of new drug approvals, compounded by liability problems faced by some major drug companies. A report by the international KPMG consultancy last month noted that, in 1998, only one out of 18 of the largest drug and medical-device makers disclosed a risk that their

pipeline of new products would not deliver. In 2003 this figure was 13.

KPMG also found that investments in pharma stocks carried more risk than other S&P 500 stocks. Their typical year-on-year change in cash flow as a percentage of assets, for example, was twice that for the market as a whole. In exchange for this risk, investors expect higher returns. But as the graph shows, they're not getting them.

Garret FitzGerald, who tracks industry trends at the University of Pennsylvania, Philadelphia, predicts that the drug sector will undergo marked structural changes in the next few years. He thinks firms able to exploit the promise of individually tailored drugs will fare best. ■

Roxanne Khamsi

Call for a cull of pointlessly different reference styles

SIR — Coping with the multitude of formats imposed by academic journals for citing references to the literature is an aggravating and labour-intensive experience. Some of this staggering profusion of styles can be reasonably rationalized, for example the financial imperatives for saving space. But, as editor-in-chief of *DNA Repair*, I find that my discussions with the managing editors and publishers of several prominent scientific journals reveal little else by way of rational decisions.

What difference can it possibly make if an author's initials are placed before or after his/her surname, or where exactly in the citation the date of a publication is situated— not to mention the myriad variations of required fonts, italics, colons, commas and full stops? And does it really make a material difference whether references are arranged by author name (alphabetically or not) or numerically, and are identified in the text by number (which may be required in superscript or not) or by author name(s)?

The prevailing attitude seems to be that we are irrevocably stuck with this state of affairs and that trying to obtain consensus among editors and publishers to adopt a universal format would be like herding mosquitoes.

This letter is by way of an appeal to the publishers and editors of major scientific periodicals to agree on a single-standard reference format. Such an initiative would, I hope, go a long way to cajoling the remaining mosquitoes to join the herd. Exactly what this format should be is outside the province of this missive. But I doubt if many would disagree that a useful and sensible method would be to list references in alphabetical order, including only the first three author names, the entire article title, the journal name, the volume number, the first page number, and the publication year.

In recent years, computer programs have been designed to facilitate the management of varying reference styles. However, these are far from perfect, and I am informed by one prominent journal that staff members are routinely obliged to correct these in order to ensure conformity to the journal's required style. Furthermore, some of these programs use a series of macros that can interfere with the typesetting programs used by publishers.

With the advent of electronic publishing, additional problems have surfaced. A consortium representing many of the leading academic publishers has established an electronic article-linking system in which articles are assigned a unique and irrevocable digital object identifier (doi). The doi consists of a unique alpha-numeric character string that is assigned to an article by the publisher

at the time of electronic publishing. Aside from the fact that there is no uniform style for citing a doi, many identifiers do not even clearly identify the journal of origin.

I find it nothing short of pathetic that the scientific community has endured this seemingly arbitrary imposition for so long!

Errol C. Friedberg

Department of Pathology, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390-9072, USA

Noting Croats' difference from other Slavs isn't racist

SIR — I am not among those who think that Dragan Primorac is a great minister of science and education, and I may have my doubts about how good a scientist he is. But your News story "Race claims spark fury over Croatia's school curriculum" (*Nature* **437**, 463; 2005) builds on tensions in this case in an unhelpful way. As a Croat, I could just as easily accept Croats being Slavs as I could their being ethnically distinct.

There is nothing racist in Primorac's claim that one genetic marker differentiates Croats from other Slavs. To call it "potentially incendiary in the Balkan region, recently torn apart by civil war" does nothing but provoke bad feeling. As a Croat and, more importantly, as a scientist, I am eager to know whether Primorac's theory is correct or not. But I will reserve judgment until there is an evidence-based scientific explanation of his interpretation of the *Science* paper, as quoted in your News story.

After all, Primorac is quoted as saying "We need much more scientific evidence before we draw conclusions", and the coordinator of the school curriculum says that examples such as this would not be included in textbooks. So I am wondering about who is actually furious, and why — and most importantly, what is the scientific rationale for being furious?

Ognjen Čulić

Medvedgradska 70, 10000 Zagreb, Croatia

Internet forest-watchers a new force for conservation

SIR — Your News in Brief story "Enthusiast uses Google to reveal Roman ruins" (*Nature* **437**, 307; 2005) made me wonder whether the time for a new conservationism is not already on the horizon. A time when the convergence of imaging and grid computing technologies, together with a policy of free access to catalogues of high-resolution Earth imagery, will make it possible to monitor selected patches of forest across the globe using the Internet.

Indeed, new instruments such as the wide-field imager to be launched with the third Chinese-Brazilian Earth Resources Satellite in 2008, with a spatial resolution of 70 metres and a revisiting time of less than a week, are ideally configured for this task.

This resource will be free of charge to users in Brazil and could be in other parts of the world too, depending on agreements with their governments. Meanwhile, a lower-resolution resource, MODIS, is already available at <http://modis.gsfc.nasa.gov>. Although rainforests are often covered by clouds, reducing the amount of time that data are available, the large numbers of potential watchers would maximize the use of data.

Today, thousands of enthusiasts are using their home computers to search for signs of extraterrestrial intelligence. Perhaps there could soon be a global network of forest-watchers, pushing the alarm button every time their protected gardens are under threat.

Fernando Manuel Ramos

National Institute for Space Research (INPE), Avenida dos Astronautas 1758, 12227-010 São José dos Campos, São Paulo, Brazil

Later results don't confirm antidepressant suicide link

SIR — Your News story "Adult suicides linked to popular antidepressant" (*Nature* **436**, 1073; 2005) reports the results of an analysis concluding that the antidepressant paroxetine is associated with suicide risk in adults and should therefore be restricted, although other studies have shown no cause for concern. But a much larger analysis by the European registration authorities (EMA) in 2004 concluded that, although suicide risk is present in children and adolescents and possibly in young adults aged 18–29, it is not present in adults above that age (see www.emea.eu.int/pdfs/human/press/pr/1120604en.pdf).

The analysis cited in your News story looked at 16 studies, comprising 1,466 patients, that were part of the original registration dossier for paroxetine in 1989. Numerous randomized clinical trials have been conducted since then, though; the EMA analysis used results from 171 randomized clinical trials, involving more than 14,000 patients, of whom about 5,000 had depression.

Although the first studies may suggest a possible elevated risk, the accumulation of evidence since then does not confirm this suggestion. Conclusions drawn from the EMA study have been integrated into the product information for paroxetine.

Tamar Wohlfarth, Jitschak Storoosum

Medicines Evaluation Board of The Netherlands, PO Box 16229, 2500 BE The Hague, The Netherlands

COMMENTARY

Deeper into the genome

The next large-scale human genome project after HapMap should catalogue inherited variation in the general population that directly affects gene function, argues **Richard Gibbs**.

The International HapMap Project has catalogued the patterns of more than 1 million single-base changes (known as single nucleotide polymorphisms, SNPs) in the genome sequences of 269 people drawn from four diverse human populations¹. Most of these SNPs do not directly influence gene function, but the data provide valuable information about the overall pattern of chromosome organization and offer new tools for finding disease-causing genes in humans.

A complementary but more direct way to discover disease-causing genes is 'medical resequencing' (MRS). Key parts of suspect genes are sequenced and compared between patients and controls to identify genetic variations that may contribute to disease². This approach has been popular and fruitful, although it relies heavily on picking the right candidate gene at the outset. MRS activity is increasing, with a few sequencing centres now analysing hundreds of individual genes. Using MRS to analyse all known human genes (currently more than 20,000) in selected sets of patients and controls is also under discussion. This would greatly enhance the chances of successful disease-gene searches.

Rare finds

Such large-scale projects favour a centralized effort, similar to that used to decipher the human genome. Here, we argue that MRS activities should be accelerated, but the goal should be to discover genetic variation in the general population (not just patients and controls) that can potentially affect gene function directly. During this sequencing phase, we would not need to know the disease status of the individuals sampled. This rich catalogue of genetic changes, here called 'functional variants', would include SNPs that alter amino acids in proteins, and possibly gene-splicing or expression levels. Most would be rarer than those pursued by HapMap. This functional-variant database would be immediately available for all



Directly analysing genes could shed light on the causes of disease.

researchers and would have considerable impact on future disease-gene studies.

Some of the functional variants that cause disease have already been catalogued thanks to studies of mendelian diseases — human conditions with simple (one-gene) inheritance³. Polymerase chain reaction (PCR), fluorescent DNA sequencing and other techniques have enabled the discovery of about 1,700 mendelian disease genes, most of which have multiple functional variants. These diseases are rare, affecting about 1 in 10,000 to 1 in 100,000 individuals, and the frequency of the mendelian functional variants in the general

population is often less than 0.05%.

The HapMap data will help to reveal functional variants that cause more common diseases, such as adult cardiac diseases, cancer and schizophrenia. These disorders mostly have complex (multi-gene) underlying genetics and the HapMap tools work best when the functional variants involved occur in the diseased population with a frequency greater than 5% (ref. 4).

The process of detecting the presence of a particular SNP is called 'genotyping' and the HapMap project has identified a subset of the estimated 3 million common SNPs that can be identified in genotyping assays. The project also identified a further subset of 'tag' SNPs that are most useful in genetic association studies, because they link common haplotypes — larger blocks of DNA that are inherited together. Using these tags, the screening of whole genomes in patient sample collections is a much more realistic proposal than it was just three years ago.

Beyond HapMap

When disease-causing functional variants fall below 5% in the diseased population, approaches aided by HapMap lose power. These limitations of HapMap are entirely expected, as the overall distribution of SNPs seen in human populations follows the shape of a power-law distribution, with the rarer SNPs

accounting for most overall variation (see left side of graph overleaf)¹. Rare functional variants corresponding to the mendelian disorders are in this category. More common SNPs appear on the right of the graph and include the few disease-causing functional variants that have been identified.

Markers that fall in between, in the frequency range 0.05–5%, are less well known, primarily because technical hurdles prevent their discovery. Nevertheless, they are likely to contribute to human disease — theoretical modelling supports their importance^{5,6} and direct evidence from disease examples is

FULI YU

emerging. Cohen *et al.* recently studied individuals who are genetically susceptible to cholesterol deposition and therefore heart disease, and who were further characterized by the mean levels of low-density lipoprotein (LDL) in their blood⁷. They reasoned that functional variants in key genes would be found most easily in the individuals with the highest and lowest levels of LDL. Using MRS, Cohen and his colleagues discovered a pair of functional variants in a key gene, *PCSK9*, that conferred low LDL levels. Notably, the functional variants they discovered to be associated with low LDL levels were in the category of 'very rare mutations', but were found at sufficiently elevated frequency in the general population to be considered 'low-frequency variation'. These functional variants would probably not have been identified using HapMap genotyping.

Bigger and better

A global search for putative functional variants provides a logical framework for extending the HapMap resource. We propose a large-scale genome project to catalogue rare genetic variation — beyond what HapMap has achieved. Some limited public and private initiatives already exist along these lines⁸, but a larger and more ambitious programme that targets diverse human populations, independently of disease or phenotype, is required to accelerate discovery of the functional variants present at low frequency.

This project would first build a functional-variant database by using MRS targeted to the coding regions of all 20,000 known human genes, plus an additional segment of the presumed promoter (gene-control) region of each. Analysis of 2,000 individual DNA samples (covering 4,000 chromosomes) would offer a good chance of finding variation that is present at the 0.05% level in the general population. This is therefore a bold proposal that could capture most of the expected functional variants (perhaps 50,000 to 100,000)⁹ found in humans.

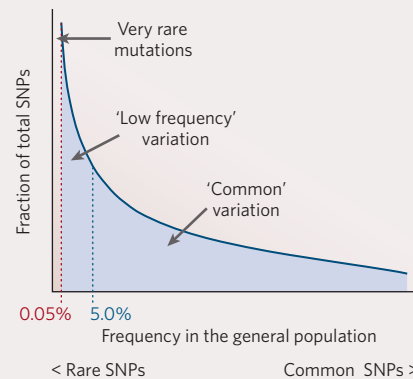
The populations sampled should include African, Asian and Caucasian representatives, in addition to Hispanic and smaller native groups. The optimal representation of different populations will be difficult to define without more data, but these choices may not be critical, as the large sample size will ensure that even large frequency differences between populations will not obscure individual variants.

The scale of this project would require improved technologies for MRS, which are already under way. MRS is easier, technically speaking, than *de novo* sequence determination because there is always a target sequence for comparison. This means that faster reductions in MRS costs are more likely than those observed for other genome projects.

Crucially, the functional-variant approach would avoid centralization of disease analyses, while exploiting the sequencing power of large genome centres to build the primary database. The overall goal would be to identify variation

GENETIC VARIATION IN HUMANS

Variation is measured by single nucleotide polymorphisms (SNPs).



that could subsequently be tested by the same kinds of genotyping methods used for HapMap. We can expect private companies to generate 'DNA probes' to test for these genetic markers in new populations, as has already occurred with the existing database of around 20,000 publicly available SNPs¹⁰. The probes could be used in a distributed fashion — in individual investigator laboratories — as part of each ongoing disease discovery effort.

The functional-variant proposal contrasts with similar ideas for large-scale MRS that focus on patient populations, such as the Cancer Genome Project¹¹, which aims to catalogue all major mutations that occur in the

"A larger and more ambitious programme is required to accelerate discovery."

most common human cancers. Such projects risk revealing patient identities, assuming that their DNA sequence data is publicly released (as occurred for the Human Genome Project).

A 2004 study showed that as few as 75 SNPs could be used to identify an anonymous patient¹². Because intense genetic analysis of disease samples requires simultaneous sequencing of several genes from each individual, identification is easy. In the functional-variant database, the free public release of the SNP data could not be used to identify the contributing individuals because only the mutations themselves would need to be fully available. The follow-up process of discovering these multiple variants in the same individual would only occur in the laboratories of individual investigators working on specific diseases, at which point patient consent could be obtained.

The functional-variant approach has other advantages over large-scale MRS projects on patient populations. Although there are thousands of well-characterized tissue samples in 'disease collections', most of these are dispersed among the laboratories of different clinical scientists. Little coordination over phenotyping

or storage exists, and collection protocols and ethical approvals are not standardized. Plans to centralize these collections, perform MRS and then publicly release the data will have to contend with complex logistics and Institutional Review Board issues. For clinicians, assigning credit for what might be years of sample collecting becomes a greater concern when all the attention is focused on new sequencing studies. The functional-variant database avoids most of these issues by separating the initial sequencing phase from the laboratory investigations of specific diseases.

One disadvantage of the Cancer Genome Project is that each new mutation discovered might be present in only a small fraction of the cells in a minute tissue sample. Non-cancer tissue from each patient must therefore be checked, to ensure that the mutation is related to the cancer and is not simply inherited rare variation. This complication does not, however, apply to the functional-variant database.

At current prices, with PCR Sanger fluorescent DNA sequencing, the cost of analysing the full complement of coding genes for 2,000 individuals would be about US\$750 million. We can, however, reasonably expect MRS costs to drop progressively, and if the work is spread over five years it is likely to cost less than \$500 million. This is more than five times the amount spent on HapMap, but is lower than the projected \$1.35 billion cost for the Cancer Genome Project.

Finally, any proposal for large-scale MRS raises a familiar dilemma in genomics — how to balance the efficient high-throughput sequencing of large genome centres with the powerful research approaches provided by the wider community. Among all possible future projects, the functional-variant proposal offers to enhance the efforts of established networks of biomedical investigators in a way that echoes the Human Genome Sequence project.

Richard Gibbs is at Baylor College of Medicine, Houston, Texas 77030, USA.

1. International HapMap Consortium *Nature* **437**, 1299–1320 (2005).
2. Gibbs, R. A. *et al.* *Genomics* **7**, 235–244 (1990).
3. Institute of Medical Genetics, Cardiff, UK www.hgmd.cf.ac.uk/hgmd0.html
4. Belmont, J. W. & Gibbs, R. A. *Am. J. Pharmacogenom.* **4**, 253–262 (2004).
5. Pritchard, J. K. & Cox, N. J. *Hum. Mol. Genet.* **11**, 2417–2423 (2004).
6. Reich, D. E. & Lander, E. S. *Trends Genet.* **17**, 502–510 (2001).
7. Cohen, J. *et al.* *Nature Genet.* **37**, 161–165 (2005).
8. www.sanger.ac.uk/genetics/exon/ and www.celera.com/celera/applera_genomics
9. Botstein, D. & Risch, N. *Nature Genet.* **33** (suppl.), 228–237 (2003).
10. www.ncbi.nlm.nih.gov/SNP/
11. Working Group on Biomedical Technology www.genome.gov/Pages/About/NACHGR/May2005NACHGRagenda/ReportoftheWorkingGrouponBiomedicalTechnology.pdf
12. Lin, Z., Owen, A. B. & Altman, R. B. *Science* **305**, 183 (2004).

Acknowledgements: Thanks to J. Belmont, D. Nelson and F. Yu for discussions and for reading the manuscript.

BOOKS & ARTS

Power for life

Did the humble mitochondrion — the powerhouse of the cell — play a key role in the evolution of life?

Power, Sex, Suicide: Mitochondria and the Meaning of Life

by Nick Lane

Oxford University Press: 2005. 320 pp.

£18.99, \$30

John F. Allen

How living things obtain, store, convert and use energy once aroused intense debate. Careers and reputations were built and destroyed in the 'ox-phos wars' about the mechanism of oxidative phosphorylation, in which electron transport is coupled to the phosphorylation of ADP, storing energy in the form of ATP. Today there is a tendency to see the problem as solved, and oxidative phosphorylation as worthy but dull. A few abstruse mechanistic details may be unresolved, but these are for obsessives, because we know, broadly, how mitochondria make ATP. It is by the chemiosmotic mechanism proposed long ago by Peter Mitchell. So mitochondria attract the molecular biologist's damning epithet 'housekeeping'. And what could be duller?

One complication is that mitochondria have genes. However, there is no cause for alarm. These genes are few, are all for housekeeping, and have survived from the symbiotic bacteria that first brought respiratory electron transfer and ATP synthesis to the otherwise fully formed eukaryotic cell. This idea, too, was once controversial. But all is now settled, and mitochondrial origins can be left to those with a taste for theoretical biology, untestable hypotheses, and the pondering of distant, one-off events.

Admittedly, mitochondria occasionally capture the headlines. They are inherited solely through the mother, so their genomes sometimes provide decisive forensic evidence and tease modern humans about their ancestry. Then there is the production of oxygen free radicals, which can promote ageing. Furthermore, reproductive cloning must overcome the irritation of mitochondrial genomes, but these can probably be replaced from the healthy cells of a donor — a third genetic parent. In Britain, the Human Fertilisation and Embryology Authority has recently given the green light to John Burn of the Institute of Human Genetics at the University of Newcastle upon Tyne to do just this. "My belief is that we are changing a battery that doesn't work for one that does...changing the mitochondria won't affect the important DNA," states Burn

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Bad for your elf? In *The Lord of the Rings*, Arwen chose love and marriage rather than immortality. But is our own immortality limited by our mitochondria?

confidently. "Mitochondria are not part of the genetic material that we consider makes us, as human beings" (*The Times*, 9 September 2005).

The central thesis of Nick Lane's book *Power, Sex, Suicide* is the antithesis of Burn's view. I'm with Lane here: mitochondria have everything to do with what makes us who we are. Lane's audacious introduction, which describes mitochondria as "clandestine rulers of the world", heralds "striking new insights into why we are here at all, whether we are alone in the universe, why we have our sense of individuality, why we should make love, where we trace our ancestral roots, why we must age and die — in short, into the meaning of life."

I have mentioned Lane's title to several colleagues, and they smiled, as I did. Provocative, yes; memorable, yes; but seriously over the top. Isn't it time popularizers of science sobered up a bit and stopped shouting for attention? Lane has a point, of course: power means the rate of doing work, or expending energy; sex refers to the odd, unexplained fact of maternal inheritance; and cell suicide is apoptosis. But this book delivers vastly more than lurid synonyms for dry, scientific terms. These three horsemen are connected, by mitochondria, and not accidentally. That still leaves "the meaning of life". I fear that sheer embarrassment will impede this book's citation in journal articles. This is unfortunate, because parts of it qualify as primary literature, by announcing at least

two major, original and testable hypotheses. I scribbled "He should publish this" in the margin, before realizing that he had.

One new hypothesis explains "why there are two sexes". Lane proposes that a tuning, or dialogue, occurs between the nuclear and mitochondrial genetic systems during the formation of egg cells. Thus the real distinction between male and female is that male gametes (sperm) have mitochondria that must be eliminated so as not to interfere with this specific, carefully selected rapport. Perhaps egg mitochondria may also be pure genetic templates, and energetically disabled, as outlined in Lane's previous book *Oxygen* (Oxford University Press, 2002). To my mind, Lane's proposal suggests that reproductive cloning, no matter how 'healthy' the mitochondria, is taking out a genetic mortgage that future generations will have to repay, and with interest, as donated mitochondria will not have been selected to be compatible with the nucleus. A further proposal is that genetic imprinting is a consequence of having tuned only half of the fertilized egg's nuclear genes: those from the mother. Another new idea is that damaged mitochondria might be replaced, by intracellular selection, during ageing.

The general reader is here forewarned that Lane comes down decisively in favour of some theories that are still regarded as 'fringe', and against others that have influential and highly cited ('respectable') proponents. I am

sure this book will be attacked elsewhere as unbalanced. Bear this in mind when you read about Mike Russell's superb scenario for the chemiosmotic properties of the earliest cells, established by geothermal convection. Consider it, too, when reading of "the hydrogen hypothesis for the first eukaryote" from Bill Martin and Miklos Müller. Believe me, people get angry at the idea of the primordial eukaryote being a methane-producer in partnership with a hydrogen-excreting anaerobe that was ancestral to both the mitochondrion and the obscure hydrogenosome. Perhaps the 'hydrogen hypothesis' is an affront to eukaryotic dignity. Personally, I think it is a liberating idea, and one that will stand the test of time.

The book was written for anyone interested in some of the most profound questions of twenty-

first-century science. The central proposals of *Power, Sex, Suicide* are clearly and forcefully propounded, are serious, have far-reaching consequences — and may even be correct. After all, not so long ago, the chemiosmotic and endosymbiont hypotheses were championed only by those thought to be mad, bad and dangerous to know. Now we read that "the dynamics of the respiratory chain are a force that has shaped the whole trajectory of life". This is a new take on why we are here. Perhaps all genes are 'housekeeping' genes, and vectorial electrons and protons were the authors of evolution — and are still its movers and shakers. Perceptions change. Do, please, read this book. ■

John F. Allen is in the School of Biological and Chemical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK.

As Daniel Charles makes clear in *Between Genius and Genocide*, his fine new biography, which will be reviewed shortly in *Nature*, Haber's ambition had a somewhat desperate quality to it. Thiessen captures this aspect of Haber's character from the very beginning, with a scene introducing Haber's relationship with fellow exile Einstein. Haber, convincingly portrayed by Aasif Mandvi, has political connections, and revels in the upper hand he holds over his less worldly counterpart, who cares little for the approval of the nation of his birth, and for whom the life of the mind is its own reward.

Haber's courtship of Clara, the philosophical tug-of-war that constitutes his relationship with Einstein, the moment of insight in his work on the production of ammonia, and the horrifying scene at Ypres are all written and staged with eloquence and flair. A climactic scene, in which Haber resigns his post rather than fire scores of Jewish scientists on his staff, is gripping. Thiessen is also aided by John McDermott's set, which, with its period chalkboards and central spiral staircase, lends an air of authenticity throughout.

Yet it must be said that a certain heavyhandedness hovers over several moments in the play. In their final meeting, Einstein's gift to the dying Haber is a tallit (a Jewish prayer shawl). Symbolically this brings the play full circle, but there is no evidence that Haber's lifelong distance from his religion caused him more pain than did his exile from Germany. The play also begins and ends with a simulated nuclear detonation. This is less subtle than it needs to be. After the final flash, Einstein wanders across the stage in anguish, wondering aloud if it is possible to go back in time. "Am I to censor all of my thoughts?" he wonders. That's an excellent and necessary question — a gift in itself. But its relevance to Haber, who was an active participant in the creation of the military-industrial complex, is not at all clear.

Despite these discordant notes, *Einstein's Gift* brings Haber's unhappy story powerfully to life. ■

Alan Packer is senior editor at *Nature Genetics*.

THEATRE

Two exiles

Einstein's Gift

by Vern Thiessen, directed by Ron Russell
Acorn Theatre, New York, until 6 November 2005

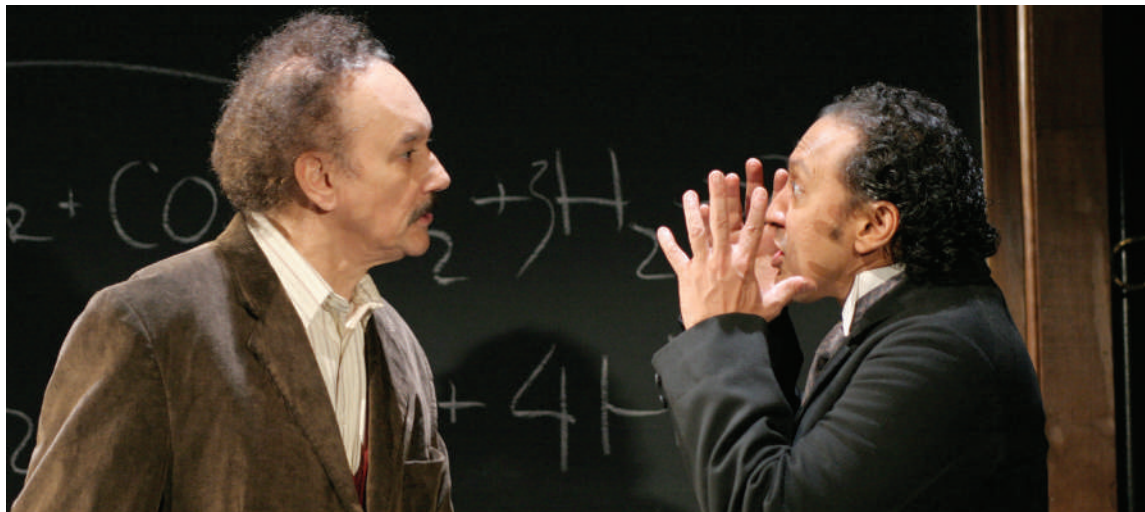
Alan Packer

If a writer wants to dramatize chemist Fritz Haber's remarkable and agonizing story, there is one sure-fire way to attract a general audience: bring Albert Einstein on stage as the narrator and build your work around the friendship between the two men. This is savvy marketing, and has the advantage of being true to life, as Einstein and Haber were indeed friends and colleagues. The tension between their two personalities is the driving force in Vern Thiessen's prize-winning play, *Einstein's Gift*, which has just received its US premiere in an absorbing and well-acted production.

It doesn't take much artistic licence to generate an air of tragedy around the facts of Haber's life. Born in 1868 to a Jewish family

in the Prussian city of Breslau, he was baptized at the age of 24. A bulldog of a man, he held, in dangerous combination, an uncritically patriotic view of Germany and a devotion to science as a means to serving it. His major scientific achievement — the Haber-Bosch process for nitrogen fixation, for which he won a Nobel prize — was immense, allowing fertilizers to yield crops that now feed billions of people.

Hoping to hasten the end of the First World War, Haber pioneered the development of chlorine gas as a battlefield weapon, and oversaw its use near the town of Ypres in Belgium. This event, with its obvious implications for warfare during the rest of the twentieth century, had personal consequences as well. It was quite possibly the trigger for his wife Clara's suicide. After the rise of the Nazi party in the 1930s, Haber, being Jewish, was cast aside by the country to which he had devoted his life. Exiled to Switzerland, depressed and in failing health, he died in a Basel hotel room in 1934.



Einstein's Gift explores the relationship between Einstein (Shawn Elliott, left) and Fritz Haber (Aasif Mandvi).

Relative beginners

It's About Time: Understanding Einstein's Relativity

N. David Mermin

Princeton University Press: 2005. 256 pp.
\$29.95, £18.95

Derek Raine

Many years ago I was asked to write a book for the 'average teenager' explaining Einstein's theories of relativity. Being naive, and not knowing any average teenagers, I took up the challenge. I thought that I had managed rather well in providing a simple account, until I met adult readers and evening-class students who told me how difficult they had found my book. Well, perhaps the subject is difficult. Hermann Bondi used to say that the public would not understand relativity until there were relativistic toys to play with. Today, computer games could play a role in teaching younger students, and lectures with animations on the web could help a more general audience. But we still turn to explanations of relativity in print.

Relativity can be summed up in a phrase that Einstein used when reflecting on the origin of the theory: "At last it came to me that time itself was suspect." It is the way that this statement is unpacked that distinguishes the many expositions of the subject. David Mermin brings to the task a lifetime of experience in making relativity accessible to the non-specialist student without simplifying more than Einstein's well-known dictum would allow.

It's About Time grew out of an earlier book for high schools and Mermin's lecture notes from a course on relativity for non-scientists at Cornell. The book begins with a thorough consideration of frames of reference. Problems involving the collisions of particles are solved by choosing a frame in which the solution is obvious (usually the centre-of-mass frame). This helps the reader to become familiar with transformations between frames of reference in a newtonian context before using them in a thoroughly unfamiliar one. The discussion also nicely illustrates the principle of relativity.

The author's starting point for the treatment of relativity is the velocity addition law. This was also a key element in Einstein's approach and not, as often treated in textbooks, a simple consequence of the Lorentz transformation. The derivation is obtained by looking at a race between a particle and a photon in various frames of reference, together with the one additional element that characterizes Einstein's relativity: the constancy of the velocity of light.

Mermin gives a central role to a quantitative statement of the relativity of simultaneity. This leads to the slowing down of moving clocks, the contraction of moving rulers, and the formula for the Doppler shift. Typical of the originality of the approach is the intriguing fact that the velocity of light, c , is 1 foot per

nanosecond, to within about 2%, which dissolves the mystery for the general reader of making $c = 1$.

The treatment is complete and involves no more than elementary algebra. Mermin suggests that his readers skip the algebraic manipulation if necessary. This is not the same as skipping the algebraic formulation of the argument, which is key to the exposition. For me the algebraic manipulation is the easy bit: relativity, with all those trains, particles and light beams, always seems to involve arguments that slide away when one tries to reproduce them. This is one reason why relativity professionals rely on space-time diagrams.

The chapter on space-time diagrams is perhaps the most important in the book. All the previous results, together with the invariance of the interval, are re-derived from this point of view, but in a way that avoids analytical geometry and uses only geometrical reasoning (essentially, similar triangles).

As the author says, one cannot write a book on relativity without including a chapter on $E = mc^2$. Mermin treats this by means of modifications to the conservation laws for momentum, as required by the relativistic addition law for velocities. The approach is quite technical, and I think that the more traditional thought experiment of a light gun

in a railway carriage provides more insight.

The final chapter contains a discussion of the reality of the Lorentz contraction (and time dilation). What causes a rod to contract from the point of view of the moving observer, or the lifetime of a muon to increase, when nothing happens to either in its rest frame? We tend to dismiss this as a question of (space-time) geometry, not physical causality, despite the fact that Einstein himself returned on several occasions to the dynamic origin of the contraction. Mermin's conclusion is that one can seek the answer in the equations of quantum electrodynamics, which are entirely consistent with the kinematic results. Is this purely a matter of taste? Any valid calculation of the length of a rod using any Lorentz invariant theory will give a result in accord with relativity — but does that mean that the calculation gives the cause of the contraction?

Mermin describes his primary audience as non-scientists, but also hopes to address undergraduate and graduate students who might find a few interesting things here. I think the problem for this secondary audience is that there are too many words, and for the general reader, too many equations. That is a pity, because this is a book full of insight with an engaging style. I recommend it to anyone who has to teach the subject to either audience: it's a brilliant basis for a set of lecture notes. ■

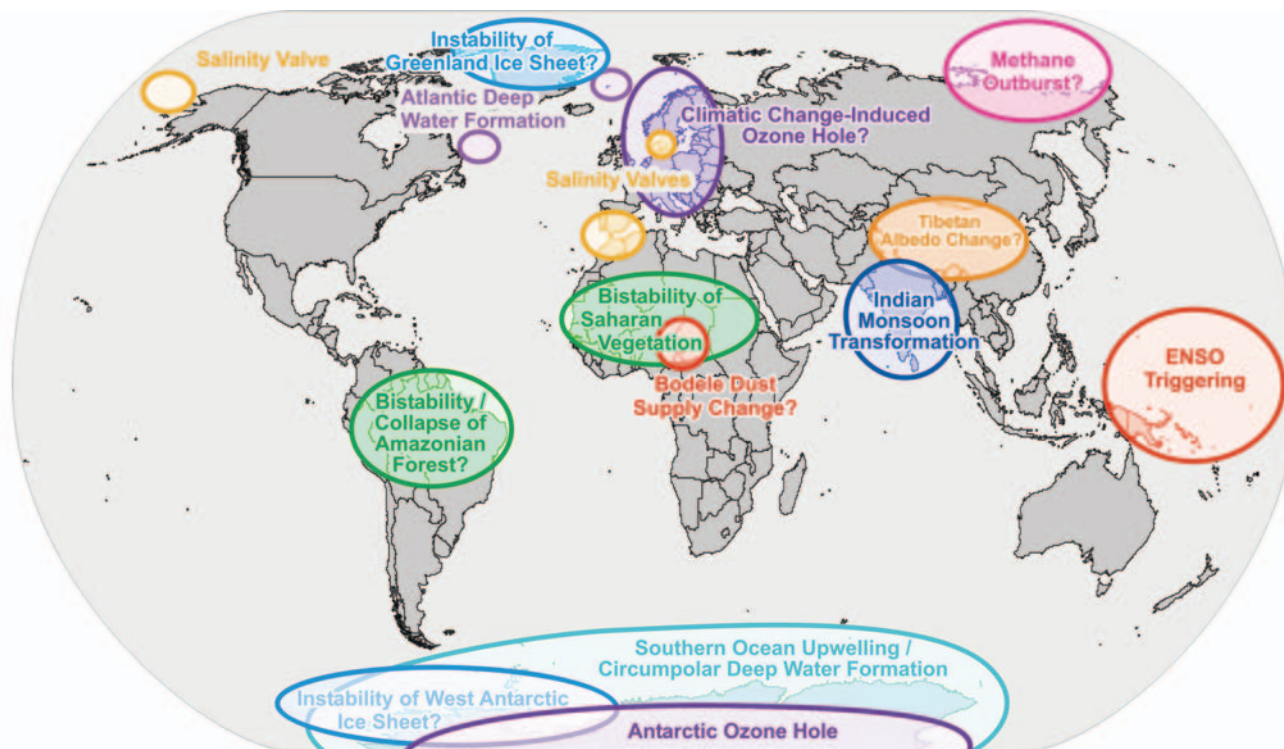
Derek Raine is in the Department of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK.

Little wonder



As a conservation biologist, Piotr Naskrecki spends his time observing some of the smaller animals that inhabit threatened biodiversity hotspots around the globe. Also an accomplished photographer, Naskrecki now shares his images of these tiny creatures, like

this armoured spider from the Solomon Islands, in his book *The Smaller Majority* (Harvard University Press, \$35, £21.95). This beautiful book is a welcome reminder that it is not just the larger, furry creatures that are under threat from habitat destruction.



Inventing an icon

Hans Joachim Schellnhuber's map of global 'tipping points' in climate change.

Martin Kemp

Any public campaign benefits from having an iconic image — something that captures the essence of the message and engraves it indelibly on our memories. But it is almost impossible to predict which images will actually stick, so creating one on demand is extraordinarily difficult. For instance, who could have forecast that of all the news photographs emanating from the Vietnam war, it was Nick Ut's photograph of a napalmed child screaming naked on a road that would become the canonical image of innocent suffering during that unhappy episode in history?

Even so, finding an iconic image was one of the goals of a meeting, 'Changing the Climate', held in Oxford, UK, on 11 and 12 September (<http://kron1.eng.ox.ac.uk/climate>). Researchers and practitioners of the visual, literary, musical and performing arts came together to publicize the predicted perils of climate change, and there was much talk about a memorable image that would encapsulate the initiative. The challenge is considerable. Any icon inevitably involves condensation and simplification, but the issues surrounding climate change are extraordinarily complex. Can an image be found that is both simple and good science? Given the contentious nature of the debates, particularly in the United States, it is unwise

to offer hostages to fortune by parading vulnerable predictions.

The data must come from the best science available, but the presentation for maximum impact is a matter of invention in art and design. Of the images produced by the scientists, one in particular seemed to have the potential to combine iconicity with complexity. This is the 'Tipping Points Map' devised by Hans Joachim Schellnhuber, director of the Potsdam Institute for Climate Impact Research in Germany and research director of the Tyndall Centre for Climate Change Research at the University of East Anglia, UK. This global map, shown here, outlines what Schellnhuber has identified as regions where the balance of particular systems has reached the critical point at which potentially irreversible change is imminent, or actually occurring.

The question marks that follow some of the labels acknowledge that it would be misleading to claim olympian certainty about the precise course of the Earth system in all respects. But what the tipping zones demonstrate vividly is the diversity and worldwide spread of climate-change threats. Some of the zonal features, such as the melting of ice and resulting rise of sea levels, have already passed the critical point and will have a literally global impact. A scientific meeting in Berlin on 5–6 October, 'Tipping

Points in the Earth System', has since focused exclusively on topics highlighted in the map.

In its present state, the map has the character of a diagram synthesizing scientific predictions. It is still short of those elusive features that make an icon, and its language remains quite specialist. This is where the visual and verbal inventors need to step in, to take Schellnhuber's neat and ingenious graphic on to the next level of accessibility and memorability. The Oxford meeting provided access to creative people with the right range of artistic skills to hone the image into a potential icon.

The map could function as a poster, as an immediately recognizable symbol for a campaign, as a memory device in all kinds of education, as the map for a series of interactive explorations on the web and other digital platforms, and as the central image for media coverage. Its graphic density in its present form dictates that it cannot function well on small scales. Perhaps a more symbolic version is needed.

Schellnhuber's map can potentially be developed into a useful image for debates about climate change. What no one can tell, however skilfully the map is refined, is whether it will achieve the levels of indelibility that will give it iconic status. Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX11PT, UK.

Exploring life's sweet spot

Glycomics: like proteins and nucleic acids, carbohydrates have essential roles in the cell, but the tools to synthesize and analyse this third class of biopolymer have, until recently, lagged far behind.

Peter H. Seeberger

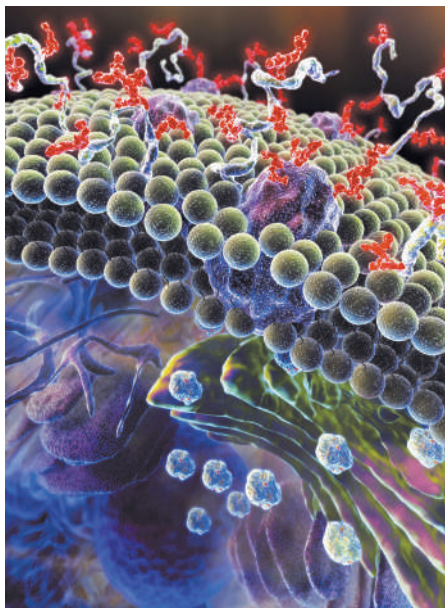
There are three main classes of repeating biopolymers: nucleic acids, proteins and carbohydrates. DNA is the blueprint of life. It contains the information that is transferred from one generation to the next and acts as a template for the synthesis of RNA. RNA in turn mediates the production of proteins — abundant and versatile molecules that form structural components and catalyse most of the reactions in living cells.

One class of proteins, glycosyltransferases, carries out the synthesis of carbohydrates, whose roles vary from energy storage and supply to mechanical support. Small oligosaccharides are linked to proteins and lipids to form glycoproteins and glycolipids, which decorate the surface of most cells and play a role in cell adhesion and recognition. But unlike for nucleic acids and proteins, which have been the main focus of biomolecular research over the past 50 years, the tools required to advance the study of carbohydrates at a rapid pace have been largely missing.

The linearity of protein and nucleic-acid polymers and the regularity of the bonds joining each monomeric unit have allowed the development of efficient and reliable tools for their analysis and synthesis. But obtaining similar tools for carbohydrates has been hampered by their structural and chemical complexity. Each monosaccharide has several free hydroxyl groups that can be used to link the unit to the next monosaccharide. This allows branching and increases the number of possible polysaccharide structures. Synthetically, carbohydrates are challenging to work with because many more functional groups have to be protected to get one specific group to react, and the stereochemistry of every new glycosidic linkage needs to be controlled.

The new millennium has seen a concerted effort to develop the tools needed for glycomics. Improvements in mass-spectrometry techniques combined with the development of enzymatic and chemical tricks have made carbohydrate sequencing faster, more reliable and more widely applicable.

Better sequencing techniques allow glyco-structures in different cell populations to be compared and, with the help of synthesis, the relevant carbohydrates to be identified. After this improved molecular understanding of specific structures (or



Sugars play a vital role in cell communication.

families of structures) is established, subsequent biological, biochemical and biophysical studies require defined molecules in much larger quantities. Our laboratory first reported the automated chemical synthesis of oligosaccharides in 2001. And although this process still cannot create all desired structures and is not yet available to the non-specialist, these goals are within reach.

Speedy access to sufficient quantities of defined carbohydrates allows the creation of tools that are commonplace in genomics and proteomics. Carbohydrate arrays that present synthetic content for high-throughput screening assays have been used to analyse interactions between carbohydrates and proteins, RNA and even whole cells. These sugar 'chips' have already yielded, for example, fingerprints of carbohydrate-binding proteins.

Rapid screening of the interactions between carbohydrates and pathogens provides insights into the role of specific carbohydrates in the infectious process, which in turn may lead to new tools for disease diagnosis and prevention. Carbohydrate affinity columns are used to isolate proteins, and *in vitro* and *in vivo* imaging using labelled oligosaccharides has become possible using synthetic carbohydrate probes. New sequencing technologies coupled to the rapid synthesis of defined oligosaccharides and their analogues have also made it possible to define carbohydrates involved in key signal-transduction events.

The glycomics advance that has perhaps had the broadest impact is that exploiting the induction of an immune response against unique pathogen or disease-specific oligosaccharides. This has led to the creation of carbohydrate-based vaccines against severe diseases including malaria, bacterial infections and cancer. Carbohydrate sequencing allows the oligosaccharide antigens to be identified and characterized, and chemical synthesis is then used to confirm antigen structure and prepare sufficient amounts of antigen to develop vaccines. Synthetic oligosaccharide vaccines against *Haemophilus influenza* type B have passed clinical trials; anticancer vaccines are currently in clinical trials; and an anti-toxin malaria vaccine is in late pre-clinical development.

With improving automated synthesis and sequencing of carbohydrates, and a host of tools for glycomics in hand, the next few years should see researchers from all areas of the life sciences drawn to this rapidly developing field, and many more biotechnology opportunities created. As well as providing the basis of synthetic vaccines against bacterial and viral infections, and cancer, synthetic glycoproteins could overcome many of the problems and limitations currently encountered when using active glycoprotein therapeutics.

Defined heparin oligosaccharides are likely to become active agents not only in heart disease but also in cancer and other diseases. Glycolipid-based signalling processes will probably be exploited for therapeutic intervention in metabolic disease and for new immunomodulators. Eventually, synthetic biology should be able to supplement chemical efforts to create large quantities of defined carbohydrates for therapeutic and diagnostic applications. And no doubt new biological roles of carbohydrates will be discovered that surpass our wildest speculations. ■

Peter H. Seeberger is at the Laboratory for Organic Chemistry, Swiss Federal Institute of Technology (ETH), Wolfgang-Pauli-Strasse 10, 8093 Zürich, Switzerland; and at the Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA.

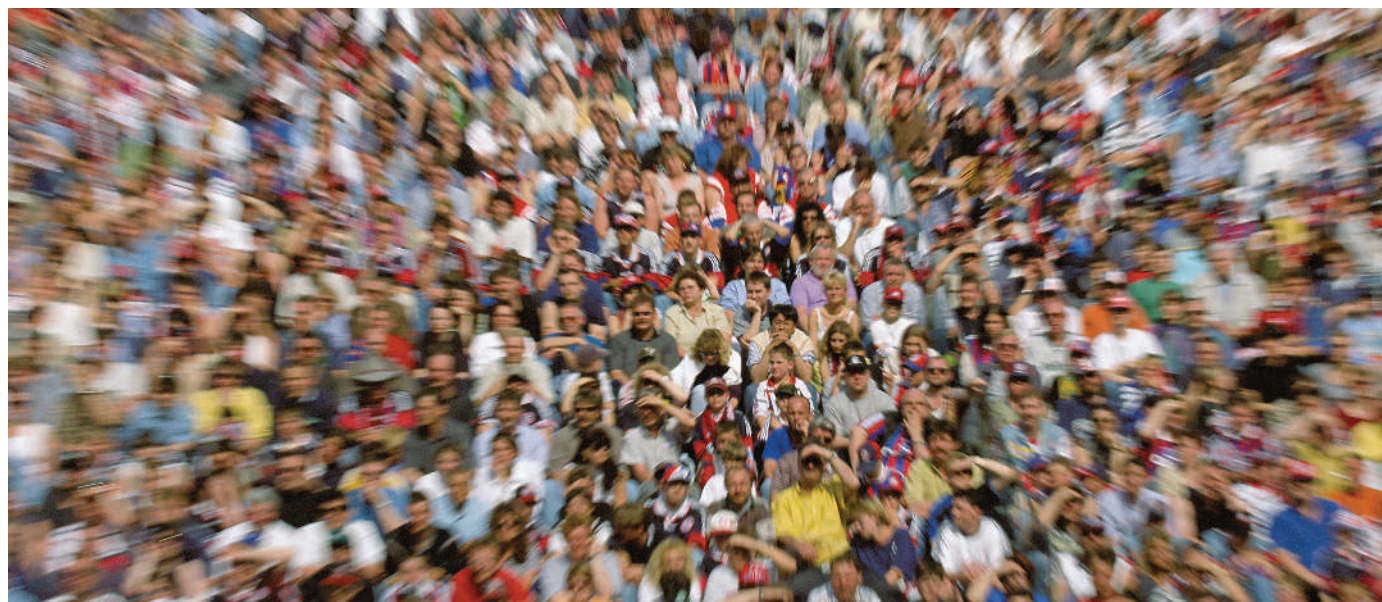
FURTHER READING

Varki, A. *et al.* (eds) *Essentials of Glycobiology* (Cold Spring Harbor Lab., New York, 1999).
Seeberger, P. H. *Chem. Comm.* 1115–1121 (2003).
The Consortium for Functional Glycomics
www.functionalglycomics.org

HYBRID MEDICAL ANIMATION/SPL

CONCEPTS

NEWS & VIEWS



GENOMICS

Understanding human diversity

David B. Goldstein and Gianpiero L. Cavalleri

The first edition of a massive catalogue of human genetic variation is now complete. The long-term task is to translate these data into an understanding of the effects of that variation on human health.

When, as a medical student, the author W. Somerset Maugham could not find a nerve where it was supposed to be, his anatomy instructor uncovered the hidden nerve and explained that “the normal is the rarest thing in the world”¹. On page 1299 of this issue², in a paper from the International HapMap Consortium, we find a detailed account of one of the primary reasons why there is no ‘normal’ human type.

The human genome has about 10 million ‘polymorphisms’, defined as genetic variants in which the minor gene forms occur at least once out of every 100 forms. Any two unrelated humans have millions of genetic differences, making them look and even behave differently. This variation is the magnificent legacy of our evolutionary past, but it comes at a price. Along with making us different in benign and interesting ways, genetics also influences health.

Modern geneticists have been hugely successful at tracking down the genetic abnormalities that lead to diseases that are inherited in a simple way in families, such as cystic fibrosis or Tay–Sachs disease. These abnormalities, however, are comparatively uncommon. The genetic disorders to which most of us will

succumb are more complex ones, such as cancer, or cardiovascular and neurodegenerative diseases. Indeed, there seems to be a grim inevitability about these common genetic diseases — as some of the ones that kill early in life are increasingly better treated, certain later killers, such as dementia, seem set to skyrocket.

The complexity of these common diseases has made them, until now, largely impervious to genetic analysis. That genetics plays an important role, however, is beyond question — as illustrated by countless studies demonstrating, for example, that genetic variation explains more than 40% of the variation of most common diseases in a population and more than 70% for some disorders such as schizophrenia^{3–5}. The difficulty is that gene variants predispose us to, but do not invariably cause, common diseases. And they predispose in combination with other gene variants and with the environment.

How do we determine which of the 10 million polymorphisms influence disease? Thankfully, it is not necessary to directly assay all 10 million sites and assess their associations with disease. This is because polymorphisms in the human genome are

often not independent of one another. When a mutation arises, it is associated with particular variants present on the same chromosome (variants that associate together are known as a ‘haplotype’). For this and other reasons, there are often strong statistical associations between polymorphisms, such that the presence of a particular variant at one site on a chromosome can predict or ‘tag’ the presence of a particular variant at another site.

The principal goal of the HapMap Consortium was to discover these associations among variants (hence the name of the project), and it has succeeded in a spectacular way. The first phase of the project, reported in this issue², consisted of compiling data on the genetic make-up, or genotype, of groups of individuals representative of four populations for more than a million single nucleotide polymorphisms, or SNPs. (SNPs are one of the most common types of variant in our genome, as opposed to, for example, insertions and deletions.) The aim was to ensure that one SNP was assayed for every 5,000 bases of sequence. The second phase, not yet complete, will considerably increase the SNP coverage.

Using a conceptually easy and reliable way to select tags, the HapMap group has

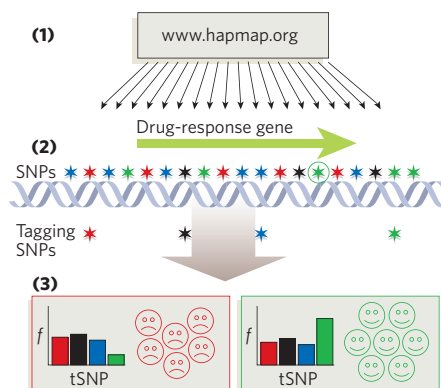
BOX 1 HapMap in practice

For any research group interested in relating human genetic variation to human health, the HapMap project is an unprecedented gift. The data make it possible to select reliable, tagging single nucleotide polymorphisms (SNPs) for genes or genomic regions in a matter of minutes.

(1) Download genotype data. Data for a gene (or region) of interest are freely available from the HapMap website.

(2) Assess the associations among SNPs in the data to select tagging SNPs. In the figure, SNPs that associate closely with each other have the same colour. The overall set of SNPs can be compressed to a subset of tagging SNPs, here depicted as one of each colour.

On average, a single SNP has between three and ten other SNPs (depending on the population) that are perfectly associated with it in the samples considered (although not necessarily in another sample from the same population). It is thus possible to greatly reduce the number of SNPs that need to be typed by excluding all but one from such sets of associated SNPs. More sophisticated approaches are likely to be even more cost-effective, and various statistical methods have been proposed. Although there is not yet a consensus on the overall best approach, one or more is



likely to emerge soon that will lead to efficient sets of tagging SNPs for the whole genome.

(3) Genetic association. The tagging SNPs are genotyped in a population sample in which individuals vary for some trait of interest, for example a good and bad reaction to a medicine. A tagging SNP that correlates with a certain reaction indicates that one of the SNPs with which it is associated influences that reaction — in this case, one of the green SNPs (circled) causes patients to have a good response to the medicine, leading to association with the relevant tagging SNP.

D.B.G. & G.L.C.

estimated that the number of tags necessary to represent common variants across the genome may be less than one-tenth of the total number of such sites. More sophisticated approaches are likely to do even better.

As well as describing these associations, the HapMap project has made other advances. Three years ago, fewer than 1.7 million polymorphisms were known. Today, thanks to the project and related efforts, that number is more than 8 million. Knowing most of the polymorphisms offers tremendous advantages. We are now in a position to apply bioinformatics tools to these data and to prioritize polymorphisms in terms of potential functionality — focusing, for example, on those that change protein sequence, or that are located in functionally important chromosomal regions (as defined by the degree of conservation of that region across species), or that seem to be relevant using other genomic criteria.

What does this mean for the study of human disease and variable response to treatment? Imagine that you wanted, before the HapMap project, to assess whether common genetic variation in the molecular target of a medicine influences patient response. You would have to re-sequence the gene in a group of individuals representative of your study population in order to identify a sufficient number of the polymorphisms. Four years ago, our group did this for a gene known as *SCN1A*, which encodes a target of antiepileptic drugs; it took us two years to identify the common polymorphisms and appropriate tags. Today, the

same job can be accomplished with simple computer algorithms, in minutes, using the HapMap data (Box 1).

The potential of whole-genome association using HapMap data is highlighted by results presented in another paper in this issue (page 1365)⁶. Cheung and colleagues restudied 27 genes whose expression had been found by family-based studies to be influenced by genetic variants. They showed that many of these would have been detected through genome-wide association, and that in one case HapMap data facilitated the identification of a variant shown by *in vitro* analysis to be responsible for altering gene expression.

The key question, then, is the relevance to human health. Because these tools are wholly new, previous failures to identify the genetic contributions to common human diseases need not presage future failures. But equally, there are no guarantees. Although we know that genetics contributes to common diseases, we do not know what sorts of gene variants are responsible. If they are common (that is, one of the 10 million), the new tool-kit will greatly accelerate the identification of disease-related genetic variation. But if the responsible variants are rarer, they will be more difficult to find. Similarly, it is not known how well other kinds of variants (for example, repetitive elements or insertions and deletions) might be represented by SNP tags. Another complexity not yet adequately addressed concerns how well the four population groups studied in the HapMap project represent variation in other human populations.

We must also admit that there is little direct evidence that the identification of risk factors for common human diseases will improve health. In those rare cases where clear risk factors have been identified, they typically have not been helpful in treatment or prevention. For example, a variant of the *APOE* gene is a strong predictor for late-onset Alzheimer's disease, but there are no known alterations in lifestyle or diet that ward off the disease. The best hope may be that risk factors will suggest new pathways for therapies, but here too there are few successful examples.

A stronger case might be made for the near-term clinical relevance of identifying genetic predictors of a patient's responses to treatment. Polymorphisms can have big effects on such responses, and identification of these effects can suggest alternative treatments. For example, there is little difference in overall effectiveness between different antipsychotic drugs, even when comparing new-generation medicines with a class introduced decades ago⁷. But there are huge differences among patients in response to the different medicines. If we could identify the gene variants that predict for example whether newer medicines cause unacceptable weight gain, or whether older ones cause a severe movement disorder, treatment options could be adjusted accordingly. The HapMap project is likely to greatly facilitate the search for such variants.

The current state of genomic sciences may be considered as a sort of awkward adolescence. The power of the modern genomic tool-kit is breathtaking. In a few years we have gone from knowing almost nothing that could be characterized as genomic (that is, focused on whole genomes rather than particular genes) to having complete genome sequences for many organisms, and now a nearly complete catalogue of the common genetic differences among people. Technical prowess is not in itself a mark of maturity in science, however. The next phase for genomics research requires a greater focus on both biological understanding and clinical utility. It is time for genomicists to turn their attention from technology to application.

David B. Goldstein and Gianpiero L. Cavalleri are at the Institute for Genome Sciences and Policy, Center for Population Genomics and Pharmacogenetics, Duke University, 103 Research Drive, DUMC Box 3471, Durham, North Carolina 27710, USA. e-mail: d.goldstein@duke.edu

1. Meyers, J. *Somerset Maugham: A Life* (Knopf, New York, 2004).
2. The International HapMap Consortium *Nature* **437**, 1299–1320 (2005); www.hapmap.org
3. Pedersen, N. L., Posner, S. F. & Gatz, M. *Am. J. Med. Genet.* **105**, 724–728 (2001).
4. Sullivan, P. F., Kendler, K. S. & Neale, M. C. *Arch. Gen. Psychiat.* **60**, 1187–1192 (2003).
5. Zdravkovic, S. et al. *J. Intern. Med.* **252**, 247–254 (2002).
6. Cheung, V. G. et al. *Nature* **437**, 1365–1369 (2005).
7. Lieberman, J. A. et al. *N. Engl. J. Med.* **353**, 1209–1223 (2005).

CHEMISTRY

A cleaner way to nylon?

Robert Mokaya and Martyn Poliakoff

The polymer nylon-6 is much in demand. An innovation in producing the precursor molecule, ϵ -caprolactam, involves a one-step process that is environmentally benign and may be scaled up for bulk production.

The catalytic acceleration of chemical reactions and improvement of their selectivity is a familiar concept. In exploiting the power of catalysis, however, chemists must often try to combine two or more catalysts to accelerate sequential stages in a complex sequence of reactions. Such a combination has great potential value, because all but the most mundane chemical transformations involve several steps. As they report in *Proceedings of the National Academy of Sciences*¹, Thomas and Raja have devised a procedure that promises to help meet this need in certain applications.

One approach to sequential catalysis involves using homogeneous catalysts, which are soluble in the reaction mixture; two or more catalysts can then be added as necessary. The catch is that it is often difficult to recover the catalysts from solution when the reactions are complete — an essential step, given that the catalyst is often more expensive than the product and must be reused many times. That problem could be solved by using heterogeneous catalysts, which are fixed to a solid support and can be filtered off at the end of the reaction. The question then becomes one of how to attach two different catalysts to the same support without compromising their activity.

Thomas and Raja¹ offer an elegant approach that exploits the properties of aluminophosphate materials. Known as AIPOs, these are nanoporous solids possessing channels that permeate the entire material, thus creating an extensive internal surface area and space that can be used to catalyse reactions or adsorb guest molecules. Nanoporous materials, particularly molecular sieves such as zeolites, which have pores of up to 2 nm in size, are widely used as heterogeneous catalysts and catalyst supports².

The properties of nanoporous solids depend on the elemental composition and arrangement of the inorganic framework, and the size of the pores. Thomas and Raja have drawn upon chemical intuition, and the principles and practices of solid-state chemistry, to design AIPO catalysts that are bifunctional in that they have both redox-active and acidic sites. The redox and acid sites are generated by the replacement (isomorphous substitution) of small amounts of the framework aluminium and phosphate with two elements chosen from cobalt, manganese or iron, and silicon or magnesium, respectively. By performing only minimal isomorphous substitution (about 4% of the aluminium atoms), they

ensured that the two types of active site were well separated (and so isolated) and uniformly distributed throughout the AIPO framework, thus generating two coexisting, single-site heterogeneous catalysts.

Thomas and Raja demonstrate the possibilities of their system by using the redox and acid sites in tandem to generate ϵ -caprolactam (CPL) from a mixture of air, ammonia and cyclohexanone. It is this choice of reaction that makes their work particularly striking. There have been previous examples of catalysts with two coexisting sites^{3–5}. But here the approach has been applied to the synthesis of an important commodity chemical — CPL is the essential monomer for making nylon-6, which has

widespread application as a textile and an industrial fibre.

Some four million tonnes of CPL are produced each year. However, the conventional manufacturing route, which proceeds via a cyclohexanone oxime intermediate (Fig. 1a), not only uses oleum ('fuming sulphuric acid') and sulphuric acid, but also is extremely wasteful. In weight terms, the major product generated by the process is actually ammonium sulphate, with CPL as a by-product. Ammonium sulphate is sold as a fertilizer, but has a limited market because it is useful only in sulphur-deficient soils, and eliminating it will be a major step forward in this chemistry. Taking another approach to that end, the Japanese company Sumitomo recently opened a small plant (capacity 60,000 tonnes per year) that uses a zeolite catalyst to generate the cyclohexanone oxime intermediate using only ammonia and hydrogen peroxide (Fig. 1b). This is followed by a new gas-phase process to convert the intermediate to CPL, with water as the only by-product.

Thomas and Raja's procedure uses neither oleum and sulphuric acid nor hydrogen peroxide. The redox sites in their AIPO catalyse the

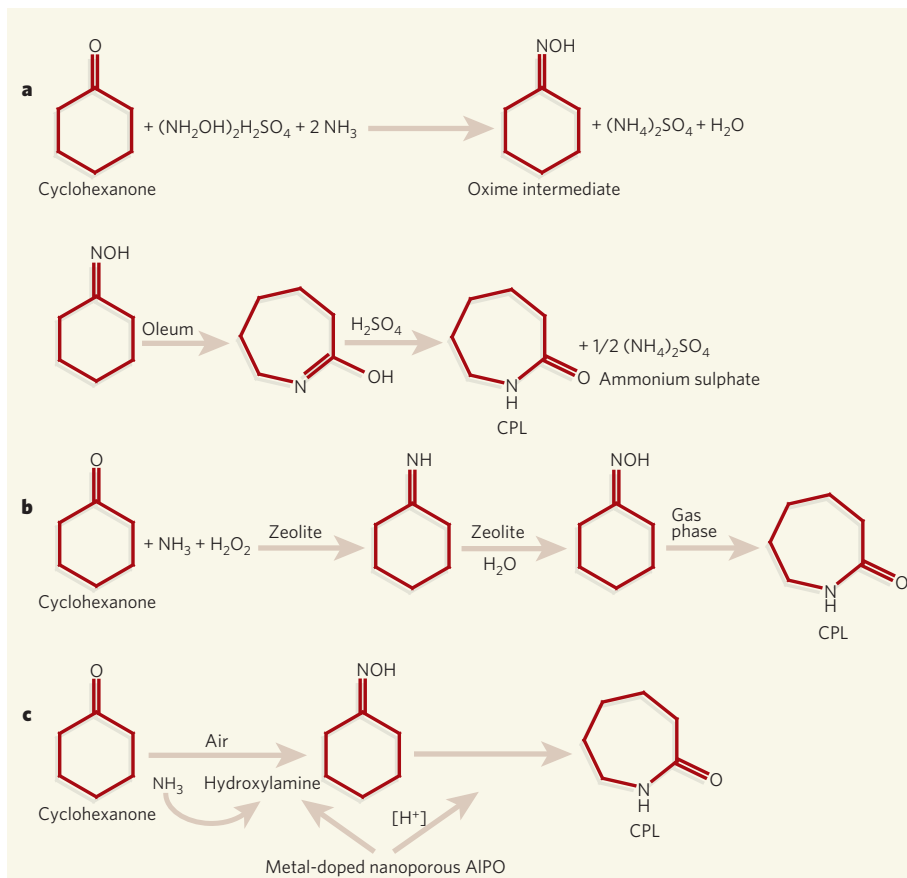


Figure 1 | Ways of making ϵ -caprolactam (CPL). **a**, The conventional route, which starts from the combination of cyclohexanone with hydroxylamine sulphate and ammonia to yield the cyclohexanone oxime intermediate. Conversion of the intermediate to CPL involves oleum and sulphuric acid, and generates large amounts of ammonium sulphate. **b**, A second approach with a zeolite catalyst uses ammonia and hydrogen peroxide as the first step in creating the oxime, which is converted to CPL in a gas-phase process. **c**, Thomas and Raja's scheme¹. Their AIPO nanoporous catalyst allows the production of hydroxylamine from air and ammonia, followed by reaction with cyclohexanone to create the oxime — which is then catalysed to CPL by the AIPO acid sites. (Reaction schemes modified from ref. 1.)

formation of hydroxylamine from ammonia and air (Fig. 1c). Reaction of hydroxylamine with cyclohexanone produces the cyclohexanone oxime, which is then converted to CPL in a rearrangement catalysed by the acid sites in the AlPO. The open structure of the catalyst (with 7.3-Å pores) allows reactants, intermediates and products to interact freely with each other and with the active sites of the catalysts. The whole reaction takes place at only 80 °C. The selectivity for CPL is respectable — up to 78% using an AlPO catalyst doped with manganese and magnesium. Such yields are still too low for a commercial process, which would require selectivity of 95% or more. But there is no reason why the selectivity should not be increased by systematically determining the most efficient catalysts to use.

The work represents real innovation in a well-studied reaction, and opens up the

possibility of applying a similar approach to many other reactions, such as epoxidation (used, for example, to produce epoxy resins and many other products). At the same time, it highlights the intellectual challenge of developing cleaner, greener chemical processes⁶. We must hope that other chemists will be inspired to rise to that challenge.

Robert Mokaya and Martyn Poliakoff are at the School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
e-mails: R.Mokaya@nottingham.ac.uk;
Martyn.Poliakoff@nottingham.ac.uk

1. Thomas, J. M. & Raja, R. *Proc. Natl Acad. Sci. USA* **102**, 13732–13736 (2005).
2. Davis, M. E. *Nature* **417**, 813–821 (2002).
3. Thomas, J. M. *et al.* *Nature* **398**, 227–230 (1999).
4. Gelman, F., Blum, J. & Avnir, D. *Angew. Chem. Int. Edn* **40**, 3647–3649 (2001).
5. Huh, S. *et al.* *Angew. Chem. Int. Edn* **44**, 1826–1830 (2005).
6. Poliakoff, M. & Anastas, P. *Nature* **413**, 257–257 (2001).

SOLID-STATE PHYSICS

Silicon's new shine

Gareth Parry

The semiconductor material used in computing systems does not emit light. But a silicon-based structure that can modulate light from an independent source might aid the marriage of optical and electronic components.

Electronic components made of silicon have so far coped with almost every new challenge in the processing and storage of data. But substantial further progress might only be possible if optical (light-based) technology is used to help distribute or communicate information within, or possibly between, processors^{1,2}. On page 1334 of this issue³, Kuo and colleagues report the first observation in a germanium–silicon semiconductor structure of an optical effect — the quantum-confined Stark effect. Their find could lead to the fabrication of silicon-based integrated circuits containing both electronic and optical components.

The optical telecommunications networks that span the globe cope with separated electronic and optical technologies. But within a compact processing or computing system, the benefits of monolithic integration — reduced cost, improved reliability and simpler packaging — are more salient. Attempts to use optical and electronic parts together in such a system, however, have always run into the problem that light-emitting optical components are made from compound semiconductors that have a very different crystal structure from the silicon so successfully used for semiconductor electronics — making integration of the two difficult.

Although silicon can absorb light, and convert it into an electrical signal, it does not itself easily give out light — unlike some other semiconductor crystals that can be used to con-

struct lasers. Silicon belongs to a class of semiconductors known as indirect-band-gap semiconductors. In such materials, light-emitting transitions between electrons in the high-energy conduction band states (which can move around the crystal and contribute to electronic conduction) and those in the lower-energy valence band states (which are localized to a particular atom) are not allowed. Over the years, attempts to get around this problem have met with little real success, and interest has shifted from the possibilities offered by silicon-based lasers to those of silicon-based optical modulators².

A modulator is analogous to a camera shutter: it transmits light when it is open, but absorbs light when it is closed. Depending on the voltage applied by the neighbouring electronic components, the modulator will either transmit or absorb a remotely generated, continuous laser beam that is directed on to it. This method of controlling light output is just as good as if the modulating material produced the light itself. The best optical modulators are based on compound semiconductor materials that use the quantum-confined Stark effect⁴, in which the allowed energy levels of electrons in very thin layers (typically 10 nanometres) change when an electric field is applied. These modulators can operate at low voltages and can be switched from an 'on' to an 'off' state at even higher frequencies than is possible using lasers.

Kuo *et al.*³ observe, for the first time, the quantum-confined Stark effect in crystalline germanium layers grown on silicon wafers. Unlike compound semiconductors, germanium (Ge) and its alloy silicon–germanium (SiGe) are perfectly acceptable materials for use in silicon electronics, as they have a similar crystal structure to that of silicon. The layers of Ge used by Kuo *et al.* are 10 nanometres thick, and ten of these are separated by SiGe layers, each 16 nanometres thick. These SiGe separating layers form a barrier to electrons and confine them in the thin layers of Ge. As a result, the energy levels of these electrons differ from the energy levels of the electrons found in the bulk crystal; the exact energy levels can be calculated quantum mechanically taking into account the shape of the well.

The thicknesses of the layers are chosen so that, when no voltage is applied, incident photons do not have sufficient energy to be absorbed (for that to happen, they would need enough energy to kick an electron in the structure into the next quantum-mechanically allowed energy level). But when a voltage is applied, the electric field changes the shape of the energy well, and the allowed energy levels of the electrons change sufficiently that the incident photons can be absorbed. Thus Ge–SiGe absorbs or transmits light according to an electronic signal from neighbouring electronic components — it acts as a modulator.

Although understood and used for 20 years in compound semiconductors for optical communication purposes⁴, this effect was not expected to be found in Ge–SiGe. This is because germanium, like silicon, is an indirect band-gap material. The ingenious technique reported in Kuo and colleagues' paper³ involves absorbing photons to higher energy levels in the conduction band of the germanium quantum wells. This takes advantage of the fact that transitions between those selected energy levels — unlike those between the lowest conduction and valence bands — are allowed. The effect that the authors observe is as strong as that seen in compound semiconductors.

Further developments must occur before we see a fully integrated silicon optoelectronic circuit. For example, most silicon electronic circuits are produced by the deposition of various oxide and metallic layers, whereas the structures used by Kuo *et al.* are produced by crystal growth. The authors are confident that their different fabrication process is suitable for mass production. Let's hope they are right.

Gareth Parry is in the Department of Physics, Imperial College London, Prince Consort Road, London SW7 2AZ, UK.
e-mail: g.parry@imperial.ac.uk

1. Cho, H., Kapur, P. & Saraswat, K. C. *J. Lightwave Technol.* **22**, 2021–2033 (2004).
2. Liu, A. *et al.* *Nature* **427**, 615–618 (2004).
3. Kuo, Y.-H. *et al.* *Nature* **437**, 1334–1336 (2005).
4. Miller, D. A. B. *et al.* *Phys. Rev. B* **32**, 1043–1060 (1985).

BIOPHYSICS

Helicase snaps back

Eckhard Jankowsky

Helicase enzymes can move along DNA or RNA, unravelling the helices as they go. But simply travelling along a nucleic acid in one direction seems not to be enough for some of these molecular motors.

Proteins of the helicase family are essential for almost all biological processes involving nucleic acids such as DNA, from the replication of the genome to the production of proteins^{1,2}. Although these enzymes are mainly known for their ability to unwind DNA or RNA helices, it has become increasingly clear that they can have many other activities. For example, helicases can displace proteins from nucleic acids and remodel nucleosomes — the protein complexes around which DNA coils^{3,4}. To account for some of the diverse activities of helicases, it has been proposed, and in a few cases shown, that these enzymes behave more like motor proteins, which move along a strand of DNA or RNA, fuelled by ATP (the cell's energy currency)⁵. On page 1321 of this issue, Myong *et al.*⁶ add an unexpected twist to the repertoire of proteins that travel along nucleic acids — the astonishingly complex movement of the bacterial helicase Rep along single-stranded DNA.

Rep from *Escherichia coli* is involved in the replication of the bacterial genome, where it helps to restart stalled replication intermediates⁷. Quite how it does this is not known. However, the DNA helicase activity of Rep is well understood, and it is clear that Rep must double up to form a dimer before it can unzip DNA^{1,8}. A single Rep molecule — a Rep monomer — cannot unwind helices, but it can move along single-stranded DNA⁸. Whether Rep exists in the cell as a dimer, a monomer, or both, is not known.

To investigate how Rep travels, Myong *et al.*⁶ used a single-molecule fluorescence technique to follow the movement of individual Rep monomers along single-stranded DNA (Fig. 1). In the presence of ATP, Rep monomers slid along DNA in the expected 3' to 5' direction. But the authors wondered what would happen if a Rep monomer encountered an obstacle on the DNA. They therefore added a second piece of DNA to form a helix in front of the monomer. A helix creates an insurmountable barrier for the Rep monomer⁸, and Rep would be expected either to fall off the DNA or to grind to a halt, remaining in front of the helix. Neither happens. Instead, Rep reaches behind itself, grabs the stretch of DNA it has just traversed, forms a loop with it, and then releases the DNA near the blockade — ending up close to the spot on the DNA where it began its journey (Fig. 1). The authors call this series of coordinated events 'snap-back', and Rep goes through many cycles of sliding forward and snapping back. Such 'shuttling'

cycles also occurred when Rep encountered other obstacles on single-stranded DNA, such as bound molecules of the protein streptavidin.

To find out how Rep snaps back, Myong *et al.* monitored changes in the shape of the protein during the shuttling cycle. Throughout the sliding leg of the cycle, Rep seems to change conformation rapidly; but as it runs into the blockade, it favours one particular conformation. The authors hypothesize that this conformation exposes or creates an additional binding site for single-stranded DNA that allows Rep also to bind the DNA over which it has just travelled. As a result, a DNA loop forms, causing Rep to release the DNA at the blockade and to snap back (Fig. 1).

Myong and colleagues also show that Rep does not require a free DNA-end to snap back. The relocation of Rep to near the end of the single-stranded DNA fragments is presumably dictated by the stiffness of the DNA strand. Although further experiments are required to discover where Rep binds to the DNA strand, and how the various DNA-binding events are coordinated, the data clearly indicate that a complex movement of a single enzyme molecule on DNA can occur in a few, relatively simple steps.

Why would it be useful for a motor to shuttle on its track? Myong *et al.* suggest that Rep might repeatedly strip off other proteins from DNA as it restarts stalled DNA replication. The ubiquitous DNA-binding protein RecA, for example, has been proposed to bind to DNA replication intermediates and so interfere with the restart of replication⁹. If true, RecA binding might need to be actively prevented, but it is not known how this is accomplished. The authors show that shuttling Rep can efficiently clear RecA from single-stranded DNA. It remains to be seen whether this process occurs in the cell. However, Rep's ability to continually displace RecA while shuttling on single-stranded DNA provides an attractive hypothesis for a physiological function for Rep.

Because the removal of proteins from nucleic acids is assumed to be a core function of enzymes of the helicase families, it will be illuminating to test whether other family members also shuttle back and forth like Rep. It will also be useful to discover whether the snap-back mechanism allows Rep or other enzymes to switch from one nucleic-acid strand to another, which could simplify the coordination of complex structural changes

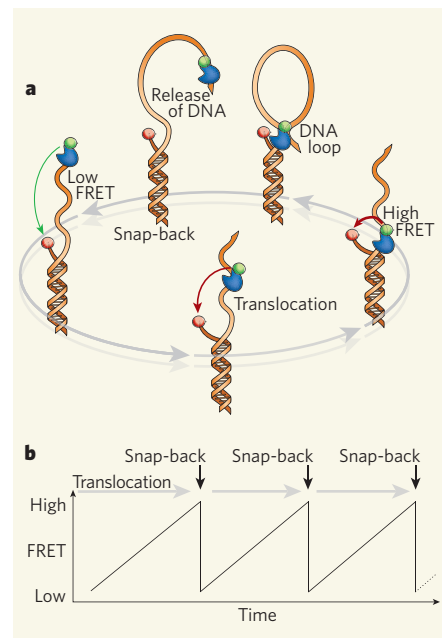


Figure 1 | Watching a helicase shuttle. **a**, Myong *et al.*⁶ followed the movement of single Rep monomers on a DNA strand by monitoring the distance between the Rep helicase and a fixed position on the DNA. They attached one fluorescent label to the DNA (red) and a second label to the helicase (green). If the fluorophores are close enough together, changes are observed in the fluorescence intensity of the two dyes, resulting from fluorescence resonance energy transfer (FRET). The FRET efficiency is highly sensitive to the distance between the two labels; low FRET shows a greater distance, whereas high FRET indicates a lesser one. Rep moves gradually along the DNA until its passage is blocked. At this point a DNA loop forms transiently, then Rep snaps back and begins its journey again. Myong *et al.* directly show the formation of the DNA loop using an alternative scheme of dye-labelling the DNA. **b**, FRET time trajectories for single Rep–DNA complexes show a characteristic sawtooth pattern, with FRET gradually increasing during the translocation phase and then suddenly decreasing during the snap-back.

during DNA or RNA transactions. Therefore, when considering possible roles for a 'helicase' in the future, we should not immediately search for the helix that the enzyme unzips, but instead remember how Rep snaps back. ■

Eckhard Jankowsky is in the Department of Biochemistry and at the Center for RNA Molecular Biology, School of Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. e-mail: exj13@case.edu

1. Lohman, T. M. & Bjornson, K. P. *Annu. Rev. Biochem.* **65**, 169–214 (1996).
2. Tanner, N. K. & Linder, P. *Mol. Cell* **8**, 251–261 (2001).
3. Linder, P. *Science* **304**, 694–695 (2004).
4. Lusser, A. & Kadonaga, J. T. *BioEssays* **25**, 1192–1200 (2003).
5. von Hippel, P. H. *Nature Struct. Mol. Biol.* **11**, 494–496 (2004).
6. Myong, S., Rasnik, I., Joo, C., Lohman, T. M. & Ha, T. *Nature* **437**, 1321–1325 (2005).
7. Marians, K. J. *Phil. Trans. R. Soc. Lond. B* **359**, 71–77 (2004).
8. Ha, T. *et al. Nature* **419**, 638–641 (2002).
9. Veaute, X. *et al. EMBO J.* **24**, 180–189 (2005).

MATERIALS SCIENCE

Changing face of the chameleon

A. Lindsay Greer and Neil Mathur

Chalcogenide materials form the basis of CD and DVD technologies. But an identity crisis looms in the wider field: what role do atomic reconfiguration, electronic processes and ionic movement play in these materials?

The chalcogens — the elements in group VI of the periodic table, particularly sulphur (S), selenium (Se) and tellurium (Te) — react with more electropositive elements, such as silver, to form chalcogenides. These are chameleon compounds: they can be crystalline or amorphous, metallic or semiconducting, and conductors of ions or electrons. Already important in optical storage discs and fibres, they are now being proposed as the basis for solid-state memory technologies. Two recent conferences — E*PCOS 05 in Cambridge*, UK, and Euromat 2005 in Prague†, Czech Republic — have demonstrated that devices using chalcogenides hinge on thermal and dynamic phenomena involving electronic, atomic and ionic processes. The links between these phenomena are not fully established, so unsuspected technological opportunities may well lie in store.

Electrical switching in chalcogenide semiconductors came to prominence in the 1960s, when the amorphous chalcogenide $\text{Te}_{48}\text{As}_{30}\text{Si}_{12}\text{Ge}_{10}$ was found¹ to display sharp, reversible transitions in electrical resistance above a threshold voltage (Fig. 1a). The switching mechanism remains unclear, but seems² to be initiated by fast^{3,4}, purely electronic processes. If current is allowed to persist in the material, it heats up, changing its atomic structure between the amorphous and crystalline states — equivalent to information being written on it. A crystalline region may be driven to become amorphous by exposure to a brief, intense heat pulse, leading to melting. The subsequent rapid withdrawal of heat sends the temperature plummeting so quickly that the melted region solidifies with the atoms still disordered. Conversely, a lower-intensity heat pulse of longer duration will crystallize an amorphous region — the crystalline state is more stable and the heat allows the atoms to mobilize just enough to assume crystalline order.

This thermally driven, amorphous–crystalline phase change, which can encode binary information, is already of great commercial significance⁵, but uses thin films of chalcogenides that are switched locally by optical rather than electrical means. In write-once and rewritable CDs and DVDs, a laser beam supplies the heat pulse for write operations,

whereas the read process exploits the relatively low optical reflectivity of amorphous data spots, or marks, in the chalcogenide film compared with that of the crystalline background. Contributions to E*PCOS 05 reporting optical-disc capacities as high as 112 gigabytes (A. Nakaoki, Sony Corp.) and data marks as fine as 20 nm (D.-P. Tsai, National Taiwan Univ.) show that current industrial trends could well continue.

Attempts¹ to induce the amorphous–crystalline transformation of chalcogenides by electrical means form the basis of phase-change random-access memory (PC-RAM). This incipient technology is on the brink of commercial exploitation by, among others, the pioneering company ECD Ovonic. For write operations, an electric current supplies the heat pulse. The read process is performed at sub-threshold voltages, and exploits the relatively large difference in electrical resistance between the amorphous and crystalline states. Recent advances in this area include a low-power device set up as a narrow line of material with the active region thermally isolated from the electrical connections, so that these are not degraded by heating⁶. There is also the prospect of much smaller data marks written with a beam of electrons⁷ and, as announced at E*PCOS 05, continuing improvements in programming times (below 30 ns: K. Attenborough, Philips Res. Leuven) and writing currents (below 0.75 mA: B. J. Kuh, Samsung Electronics).

Although the electronic transitions and atomic rearrangements relevant to both optical discs and PC-RAM featured strongly at E*PCOS, contributions from ions were not considered — even though amorphous chalcogenides can have significant ionic conductivities. At Euromat 2005, however, it was shown that ionic transport can be useful for data storage in a solid chalcogenide electrolyte. At the nanoscale, this electrolyte consists of crystalline metallic islands of silver selenide (Ag_2Se) dispersed in an amorphous semiconducting matrix of germanium selenide (Ge_2Se_3) (M. N. Kozicki, Arizona State Univ.). The performance of a prototype electrolytic chalcogenide cell is described in Figure 1b.

Technologies exploiting phase-change and electrolytic chalcogenide devices are evolving convergently. Although the microscopic mechanisms seem rather different, the two technologies display similar measurable characteristics. The sudden onset of conduc-

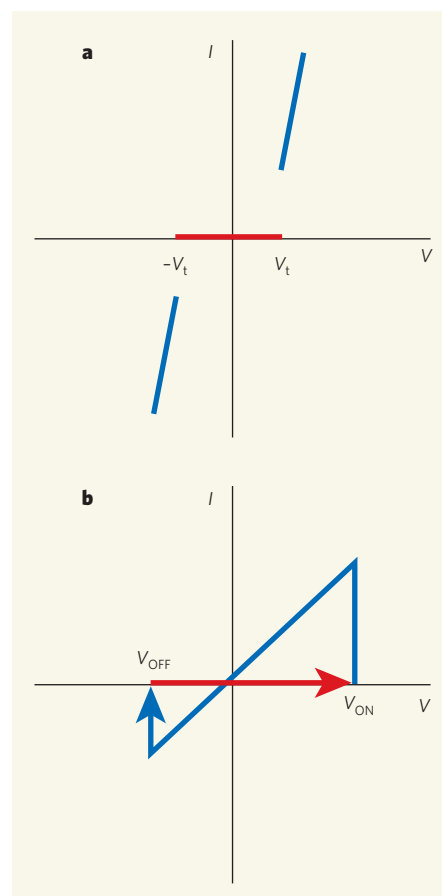


Figure 1 | Two types of chalcogenide device. **a**, Electronic switching associated with phase changes. At moderate frequencies, the alternating current–voltage (I – V) characteristic of a suitable amorphous chalcogenide material shows symmetric switching above a threshold voltage V_t from a highly insulating regime (red) to a more conducting regime (blue, gradient reduced for clarity). (Modified from ref. 1.) Optically addressed phase-change materials form the basis for CD and DVD technologies. **b**, Ionic (electrolytic) switching. A solid electrolyte of two chalcogenide phases can also encode information. In the cycle shown, the insulating state (red) switches at V_{ON} to a conducting state (blue) when nanoscale bridges of silver form between electronically conducting islands in an ionically conducting matrix. These bridges persist until they are dissolved at a sufficiently negative voltage, V_{OFF} . The asymmetric I – V characteristic shown here is reminiscent of symmetric I – V characteristics in phase-change random-access memory, where read operations in a working device would also be performed at low voltages. (Modified from ref. 10.)

tion — at V_t in Fig. 1a and V_{ON} in Fig. 1b — is, for example, associated with the formation of filamentary conducting pathways in both cases. In PC-RAM this onset is thought to arise electronically via the injection of charge carriers², but stems in the electrolytic device from the deposition of metal atoms (electrodeposition) to form conducting bridges between the islands. Another example of the similarity in characteristics but variation in underlying processes of the two types of

*European Symposium on Phase Change and Ovonic Science, Cambridge, 3–6 September 2005; www.epcos.org

†Symposium on Memory Storage Materials, European Congress on Advanced Materials and Processes, 5–8 September 2005; www.euromat2005.fems.org/index.htm

chalcogenide device is that both show incremental changes in their structure, cumulative over time, when they are operated below their threshold voltages. These changes give rise to controllable intermediate conductivities and are in effect precursors to the binary memory effects that make chalcogenides useful as storage materials. In PC-RAM, this cumulative behaviour is readily explained by crystal growth; in the electrolytic variant, it is explained by electrodeposition.

Both chalcogenide technologies present exciting opportunities that are not restricted to memory, but include cognitive computing^{5,8} (E*PCOS 05: S. R. Ovshinsky, ECD Ovonic) and reconfigurable logic circuits⁹. It is too early to tell which technology will be selected for which niche, but scientific interest alone should motivate a closer look at chalcogenide materials to investigate correlations between phase-change and electrolytic behaviour. To take one example, the migration of dissolved ions is required in the electrolytic case, but could degrade the performance of a phase-change device. Fluxes of both electrons and

ions participate in electromigration — widely studied as a degradation mechanism of the electrically conducting lines for integrated circuits. Thus, a unified approach to the study of chalcogenides, assessing the roles of atoms, ions and electrons, may prove crucial for both device performance and reliability. ■

A. Lindsay Greer and Neil Mathur are in the Department of Materials Science and Metallurgy, University of Cambridge, New Museums Site, Pembroke Street, Cambridge CB2 3QZ, UK. e-mails: alg13@cam.ac.uk; ndm12@cus.cam.ac.uk

1. Ovshinsky, S. R. *Phys. Rev. Lett.* **21**, 1450–1453 (1968).
2. Mott, N. F. in *Nobel Lectures, Physics 1971–1980* (ed. Lundqvist, S.) 403–413 (World Scientific, Singapore, 1992); <http://nobelprize.org/physics/laureates/1977/mott-lecture.pdf>
3. Adler, D. et al. *J. Appl. Phys.* **51**, 3289–3309 (1980).
4. Vezzoli, G. C., Walsh, P. J. & Doremus, L. W. *J. Non-Cryst. Solids* **18**, 333–375 (1975).
5. Strand, D. J. *Optoelectron. Adv. Mater.* **7**, 1679–1690 (2005).
6. Lankhorst, M. H. R. et al. *Nature Mater.* **4**, 347–352 (2005).
7. Gibson, G. A. et al. *J. Appl. Phys. Lett.* **86**, 051902 (2005).
8. Ovshinsky, S. R. *Jpn J. Appl. Phys. Pt 1* **43** (7B), 4695–4699 (2004).
9. Terabe, K. et al. *Nature* **433**, 47–50 (2005).
10. Kozicki, M. N., Park, M. & Mitkova, M. *IEEE Trans. Nanotechnol.* **4**, 331–338 (2005).

CELL BIOLOGY

Helices sculpt membrane

Guillaume Drin and Bruno Antonny

Many proteins are carried within cells in bubble-like sacs. These are pinched off from membranes inside the cell, and it seems that the Sar1p protein is key in both starting and finishing this budding process.

The cell contains a network of membrane-bound compartments that exchange proteins with each other and with the cell surface thanks to several haulage systems, each providing a specific link between one station and another. At the departure point, specialized 'coat' proteins wrap up a small area of the lipid membrane, shaping it into a bulging 'bud' and gathering up proteins due to be transported inside it¹. The bud detaches from the membrane — a stage called fission — to form a bubble-like 'vesicle' loaded with cargo. Lee et al.² report in *Cell* that a coat protein called Sar1p, whose structure contains several α -helices, initiates buds for one type of vesicle by thrusting one of its helices into the membrane, causing it to balloon outwards².

We knew that Sar1p begins the formation of so-called COPII vesicles, but quite how was unclear. These vesicles transfer proteins from a membrane-bound structure called the endoplasmic reticulum, where they are made, to another such structure, the Golgi apparatus, where they are processed into their final form. A common cellular fuel called guanosine triphosphate (GTP) activates Sar1p. When Sar1p binds to GTP, it exposes a short α -helix

at its amino (or N) terminus that anchors the protein to the membrane of the endoplasmic reticulum. There, Sar1p recruits two large COPII protein complexes, Sec23/24p and Sec13/31p, which polymerize into a curved lattice. Studies using artificial lipid vesicles called liposomes show that adding all these components is sufficient to generate coated buds on the liposome and, less efficiently, free coated vesicles³. Now Lee et al.² report how Sar1p contributes to the initial moulding of the membrane and, less expectedly, to membrane fission.

Using electron microscopy, the authors first show that Sar1p alone can deform liposomes into long, narrow tubules, but only when it is bound to GTP, suggesting an involvement of the N-terminal helix. To demonstrate this, they swap this helix for a peptide that binds to an artificial lipid. As expected, the Sar1p mutant still binds to liposomes containing the artificial lipid but no longer deforms them.

Both normal Sar1p and the domain-swapped mutant can interact with the other COPII complexes, so the next step was to compare incubations conducted with the complete set of COPII proteins. Puzzlingly, though, buds do form in the presence of the



50 YEARS AGO

For some time past, the B.B.C. Research Department has been studying the technique of colour television, and recently a programme of experimental transmissions was started outside normal broadcasting hours. On October 20, Sir Harold Bishop, director of technical services, presented a special demonstration for the Press. This comprised the transmission over a closed circuit at the Alexandra Palace station, of still pictures, a short travel film and a number of 'live' camera shots, all of which were reproduced at the receiving end as attractive colour pictures. From *Nature* 29 October 1955.

100 YEARS AGO

The Far East. By Archibald Little. — Of late years the Far East is only far in actual distance; it is very near to our thoughts, while the ignorance regarding these lands is being very rapidly dispelled... China stands now at the parting of the ways; for many years resolute in keeping out foreign inventions so distasteful to the old-fashioned mandarin, circumstances have proved too strong, and railways, the precursors of western life, are now being built or projected throughout the land... Consider the Yangtse Valley... This magnificent river will undoubtedly remain the great high road for commerce into Central China; but railways are and will be built to act as feeders to the main line, much to the profit of the shareholders and of the inhabitants, for Chinese are born traders, and already make use of the pioneer of Chinese railways — the line from Tientsin to Peking — in large numbers.

Finally, we have a vivid description of the southern basin, Canton, Hong Kong, and the provinces bordering on French territory. Yunnan, which adjoins our Burma, has a particular interest to Englishmen; but here, owing to our supineness in days gone by, we have allowed the French to get ahead of us with their railway, which will undoubtedly draw to itself all that is valuable of the trade of the province. From *Nature* 26 October 1905.

50 & 100 YEARS AGO

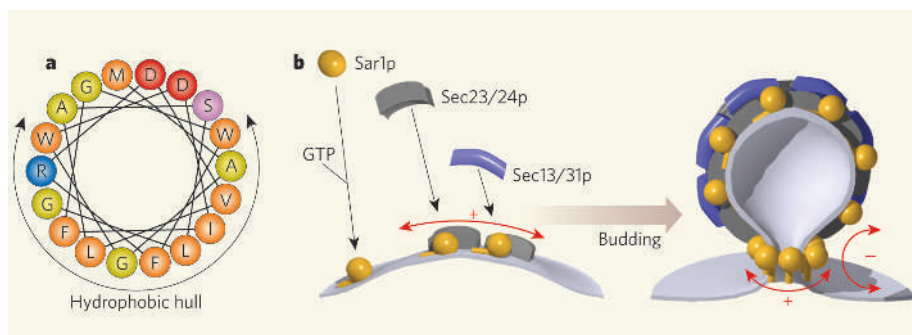


Figure 1 | Sar1p in budding and fission. **a**, The N-terminal 18 amino acids of Sar1p form an amphipathic helix, seen here as though looking along the central axis with each amino acid identified by a one-letter code. The highly hydrophobic amino acids (orange) constitute a wide side, whereas polar amino acids (red, negatively charged; blue, positively charged; purple, hydroxylated) make up a smaller one. Other amino acids are in yellow. **b**, On activation by GTP, Sar1p inserts its N-terminal helix into the membrane. Lee *et al.*² show that, once there, it bends the membrane outwards (+) and recruits the Sec23/24p complex, which polymerizes into a coat with Sec13/31p. Eventually, the coated bud is attached to the membrane by a small neck, with a negative curvature (−) in one direction and a positive curvature (+) in the other. Several Sar1p N-terminal helices may orientate parallel to the main neck axis, where they might further constrict the membrane and help fission and release of the bud.

swapped Sar1p mutant, yet free vesicles are scarce. In line with this, follow-up experiments conducted on membranes derived from endoplasmic reticulum show that an intact N terminus in Sar1p is key to the efficient release of COPII vesicles. So, if there is no doubt that the spherical shell formed by Sec23/24p and Sec13/31p is central to the sculpting of the membrane, Lee and colleagues' study² implies that the N terminus of Sar1p is not merely a simple piece of tape that sticks the COPII coat to the membrane, but that it has an active role in membrane deformation and fission.

The N-terminal helix of Sar1p is amphipathic — that is, it has a hydrophobic face and a hydrophilic face. The wide hydrophobic 'hull' should insert between the lipid acyl chains of the membrane, while the polar hydrophilic side interacts with the lipid heads and the watery environment of the cytoplasm (Fig. 1a). From model studies, we know that this kind of helix is designed to float on

biological membranes, with the axis lying at the interface between the polar and nonpolar lipid regions⁴. The membrane is a tightly packed bilayer of lipids, so when the N-terminal helices from numerous Sar1p proteins adsorb on its surface, they will expand the outer layer and, because the bilayer has a finite area, compress the inner layer. As a result, the membrane will bend and dome. Indeed, when Lee *et al.* replaced bulky amino acids in the hydrophobic side of the helix with smaller ones, Sar1p was less able to make tubules from the liposomes and to generate transport vesicles from isolated membranes *in vitro*.

Because Sar1p recruits Sec23/24p, which has a three-dimensional structure that is adapted to a convex surface, it is easy to imagine how the two proteins work in concert to bend the membrane at early stages of coat assembly⁵ (Fig. 1b). However, the role of Sar1p at the fission step is less intuitive. The curvature of the bud neck resembles that of a horse saddle,

being negative in one direction and positive in the other. If the N-terminal helix of Sar1p invades the neck, its most plausible orientation is to align along the neck axis (Fig. 1b). A ring of parallel helices emerging from the coat edge may further constrict the neck and help membrane fission. Notably, COPII-coated buds on liposomes show a wider neck with N-terminal Sar1p mutants than with the unmutated form (Figs 3 and 8 in ref. 2).

The formation of clathrin-coated vesicles, which transport cargoes from the cell surface, follows an analogous process to that of COPII vesicles in that a short N-terminal helix of the protein epsin allows the plasma membrane to deform⁶. However, the epsin helix is shorter and has a smaller hydrophobic hull. Moreover, its polar side contains several electrically charged residues that bind specifically to PIP₂, a negatively charged lipid that is a hallmark of the plasma membrane. So if the insertion of hydrophobic residues from amphipathic helices seems to be a common mechanism for inducing membrane curvature, subtle changes in the sequence may govern the ability to deform specific cellular membranes.

Hydrophobic and polar residues form the two broad classes of the amino-acid alphabet, and their segregation is the basis of the amphipathic helix. Yet hydrophobic amino acids vary in size, and this should influence the helix 'footprint' on the membrane. Likewise, the polar amino acids (such as hydroxylated, basic or acidic residues) in the other side do not interact to the same extent with the lipid polar heads^{7,8}. No doubt, the language of membrane-deforming helices at the complex membrane–water interface is very rich and remains to be translated.

Guillaume Drin and Bruno Antonny are at the CNRS Institut de Pharmacologie Moléculaire et Cellulaire et Université de Nice, Sophia Antipolis, 06560 Valbonne, France.
e-mail: antonny@ipmc.cnrs.fr

MYCOLOGY

The whiff of danger

You don't take the death cap (*Amanita phalloides*) home for tea. This species, pictured here, is infamously poisonous, with many other mushrooms being toxic to a greater or lesser degree.

Thomas N. Sherratt, David M. Wilkinson and Roderick S. Bain have addressed two issues raised by the existence of poisonous mushrooms (*Am. Nat.* doi:10.1086/497399). The first question was what purposes possession of poisons might serve in mushrooms. One possibility is that toxins are simply a metabolic by-product. Another that has

been suggested by several authors is that they act as a deterrent to predators, which might otherwise destroy the mushroom before its spores have matured and dispersed. Fungus-loving vertebrates could in particular be highly destructive.

An evolutionary principle is that if you as an organism go to the bother of being unpalatable, you might as well signal that fact. Does this apply in mushrooms? To investigate this second issue, Sherratt *et al.* turned to data compilation and neural-network analysis. They made use of

modern evolutionary trees to judge the incidence of poisonousness in mushrooms, then analysed data sets, culled from field guides, to see whether poisonous species tend to have particular ecological correlates — whether, for instance, they are more colourful, more aggregated or have a more noticeable odour.

Overall odour (and not cap colour) came out as the best predictor of toxicity, a result that was supported by pairwise comparisons of related poisonous and edible forms. Given that many animals forage by night, and that nocturnal mammals tend to have relatively poor colour vision, the authors suspect that odour provides the more effective signal.



Sherratt *et al.* make plain that their study is correlative only, and that — for them and others — this is a work in progress. There is rich scope for further investigation of the hypothesis that poisonous mushrooms use odours as warning signals, and of the likely exceptions.
Tim Lincoln

G. MCCARTHY / NATUREPL.COM

- McMahon, H. T. & Mills, I. G. *Curr. Opin. Cell Biol.* **16**, 379–391 (2004).
- Lee, M. C. S. *et al. Cell* **122**, 605–617 (2005).
- Matsuoka, K. *et al. Cell* **93**, 263–275 (1998).
- Hristova, K. *et al. J. Mol. Biol.* **290**, 99–117 (1999).
- Bi, X. *et al. Nature* **419**, 271–277 (2002).
- Ford, M. G. *et al. Nature* **419**, 361–366 (2002).
- Segrest, J. P. *et al. J. Lipid Res.* **33**, 141–166 (1992).
- Bigay, J., Casella, J. F., Drin, G., Mesmin, B. & Antony, B. *EMBO J.* **24**, 2244–2253 (2005).

SOLID-STATE PHYSICS

Spin in the slow lane

Bart van Wees

Electrons were until recently thought to transport their charge and spin equally freely through metals and semiconductors. Now it seems that spin can lag considerably behind charge.

The recognition that electrons can transport through a solid not just charge, but also spin (the intrinsic 'rotation' of electrons, which gives rise to their magnetic moment), sowed the now burgeoning field of spintronics. As charge and spin are properties of individual electrons, it is tempting to conclude that both will be transported equally efficiently. But Weber *et al.*¹, writing on page 1330 of this issue, show that the diffusion of electron spin can be substantially slower than that of electron charge. This could have important implications for spin-based electronics.

The ability to generate, study and use phenomena such as spin injection, spin currents and spin accumulation, in metals as well as semiconductors, has made it imperative to discover how similar — or different — spin and charge transport are. In metals, no substantial differences have been observed²; but in semiconductors, experiments in the past decade have already indicated that charge and spin diffusion act differently³. The comparison of the two transport mechanisms is simplified by assuming that the electron spin (or its associated magnetic moment) can point only up or down along a specific axis. In conventional charge currents, these spin-up and spin-down electrons move in the same direction. But when, in the absence of a charge current, spin-up electrons move in one direction, and an equal number of spin-down electrons move in the opposite direction, a spin current is established, with magnetic moment effectively being transported in one direction.

Usually, electron motion is diffusive: scattering on impurities, lattice defects and lattice vibrations will, after some time, randomize the direction of an electron's movement. The more scattering that takes place, the slower the electron diffusion will be. As this effect applies equally to spin-up and spin-down electrons, the naive expectation is that the effective rates of charge and spin diffusion should be the same. But this neglects the electrostatic repulsion, or Coulomb interaction, that exists between electrons. The Coulomb interaction does not affect the diffusion of charge, as it does not change the total momentum of the

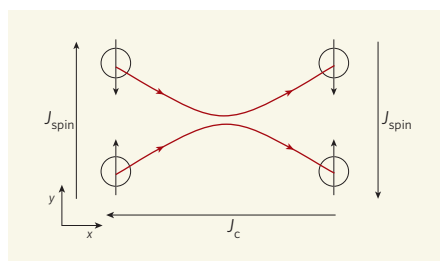


Figure 1 | Spin drag. Two electrons are moving in the x - y plane with opposite directions of spin. In this particular case, they initially contribute to a charge current J_c in the negative x -direction (conventionally, electric current is depicted as a flow of positive charges, in the opposite direction to the actual electron flow). Because the spin-up electron is moving upwards and the spin-down electron downwards, there is also a net spin-up current J_{spin} in the positive y -direction. Because of Coulomb repulsion as the electrons near each other, and the resulting exchange of momentum, their direction of motion in the y -direction, and so their contribution to the spin current, is reversed. Their contribution to the charge current, however, is unaffected. Thus, unlike the diffusion of charge, the diffusion of spin is slowed down — the spin Coulomb drag demonstrated by Weber *et al.*¹. (Modified from ref. 1.)

electrons. But it does reduce the diffusion rate of spin through scattering between spin-up and spin-down electrons (Fig. 1) — the effect known as spin Coulomb drag that has now been demonstrated experimentally by Weber and colleagues¹.

Measuring the relevant experimental parameter, the spin diffusion constant D_s , necessitated a clever technique based on a so-called spin grating⁴. The authors first used¹ the interference between two laser beams polarized linearly, but at right angles to each other, to 'pump' their sample — a crystal of the semiconductor gallium arsenide — with a very short light pulse. The two beams met in the plane of the semiconductor at an angle, so their relative phase changed linearly along its surface. This induces a periodic variation in spin density — the spin grating — in the quantum wells in which the electrons are confined. Here, regions of excess spin-up and excess

spin-down electrons alternate with a spatial period L that is dependent on the angle of incidence of the two beams. This spin grating decays with time, partially through intrinsic relaxation of the spin, and partially as a result of diffusion between the excess spin-up and spin-down regions (on a typical timescale $T = L^2/D_s$, given by the laws of diffusion).

The authors then applied a further linearly polarized, 'probe' laser pulse to their crystal. This brought the Faraday effect into play, which dictates that the interaction of a beam of light with a magnetic field — in this case, that induced by the electron spins — will bring about a (slight) rotation of the polarization of the reflected beam. The angle of rotation depends on the induced magnetization (and thus the spin density), so the decay of the spin grating can be investigated by varying the delay between the pump and probe pulses. By ascertaining the effective decay time as a function of the spatial period L , the authors determined the relative importance of diffusion and relaxation, and obtained values for both D_s and the intrinsic spin relaxation time. They found that the rate of spin diffusion indicated by D_s was considerably slower than the charge diffusion rates obtained from conventional electronic measurements.

In theoretical calculations^{5,6}, spin Coulomb drag is expressed by a parameter known as the spin-drag transresistivity; this relates the current in each of the spin channels (spin up or spin down) to the effective electric field in the opposite spin channel. The significance of the transresistivity depends on the number of spatial dimensions considered, the strength of the various interactions between the electrons, and the conventional electronic resistivity. If this last quantity is small, as it is in good conductors such as metals, the effect of spin drag will be relatively small². In the two-dimensional system used by Weber *et al.*¹, however, the effect of spin Coulomb drag turns out to be considerable. Reducing further to a one-dimensional system has revealed that spin and charge can even separate⁷, with the two intrinsic electron properties developing their own distinctive transport modes. These findings are bound to stimulate further research. What is already abundantly clear, however, is that spin and charge both move at their own pace. ■

Bart van Wees is in the Department of Applied Physics and the Materials Science Centre, Groningen University, Nijenborgh 4.13, 9747 AG Groningen, The Netherlands. e-mail: b.j.van.wees@rug.nl

- Weber, C. P. *et al. Nature* **437**, 1330–1333 (2005).
- Jedema, F. J., Heersche, H. B., Filip, A. T., Baselmans, J. J. A. & van Wees, B. J. *Nature* **416**, 713–716 (2002).
- Kikkawa, J. M. & Awschalom, D. D. *Nature* **397**, 139–141 (1999).
- Cameron, A. R., Riblet, P. & Miller, A. *Phys. Rev. Lett.* **76**, 4793–4796 (1996).
- D'Amico, I. & Vignale, G. *Phys. Rev. B* **68**, 045307 (2003).
- Flensberg, K., Jensen, T. S. & Mortensen, N. A. *Phys. Rev. B* **64**, 245308 (2001).
- Auslaender, O. M. *et al. Science* **308**, 88–92 (2005).

BRIEF COMMUNICATIONS

Darwin and Einstein correspondence patterns

These scientists prioritized their replies to letters in the same way that people rate their e-mails today.

In an era when letters were the main means of exchanging scientific ideas and results, Charles Darwin (1809–82) and Albert Einstein (1879–1955) were notably prolific correspondents. But did their patterns of communication differ from those associated with the instant-access e-mail of modern times? Here we show that, although the means have changed, the communication dynamics have not: Darwin's and Einstein's patterns of correspondence and today's electronic exchanges follow the same scaling laws. However, the response times of their surface-mail communication is described by a different scaling exponent from e-mail communication, providing evidence for a new class of phenomena in human dynamics.

During their lifetimes, Darwin sent at least 7,591 letters and received 6,530; Einstein sent more than 14,500 and received more than 16,200. We start from a record containing the sender, recipient and the date of each letter^{1,2} sent or received by the two scientists. Their correspondence exploded after their rise to fame, and reached a highly fluctuating pattern afterwards (Fig. 1a). Although, on average, they wrote 0.59 (Darwin) and 1.02 (Einstein) letters a day during the last 30 years of their lives, these

averages hide significant daily fluctuations. For example, Darwin wrote 12 letters on New Year's Day in 1874 and Einstein received 120 letters on 14 March 1949, his 70th birthday.

The response time, τ , represents the time interval between the date a letter was received and the date that the reply was sent. As shown in Fig. 1b,c, the probability that a letter will be replied to in τ days is well approximated by a power law, $P(\tau) \approx \tau^{-\alpha}$, where $\alpha = 3/2$. The fact that the scaling spans close to four orders of magnitude, from days to years, indicates that most responses (53% for Einstein, 63% for Darwin) were sent within less than ten days.

In some cases, however, the correspondence was stalled for months or years. Some of these represent long breaks in the correspondence and a few are a consequence of missing letters. Others, however, correspond to genuine delays, like Einstein's response on 14 October 1921 to Ralph De Laer Kronig's letter of 26 September 1920, which starts with: "In the course of eating myself through a mountain of correspondence I find your interesting letter from September of last year."

To understand the origin of the observed scaling behaviour, we have to realize that, given the wide range of response times, both Darwin and Einstein must have prioritized correspondence in need of a response. Thus, a simple model of their correspondence assumes that letters arrive at a rate λ and are answered at a rate μ . Each letter is assigned a priority, with high-priority letters being answered soon after their arrival, and others having to wait.

The waiting-time distribution of this simple model³ follows⁴ $P(\tau) \approx \tau^{-3/2} \exp(-\tau/\tau_0)$, which predicts a power-law waiting time for the critical regime $\lambda = \mu$, when $\tau_0 = \infty$. Given that Darwin and Einstein answered only a fraction of letters they received (their overall response rate being 0.32 and 0.24, respectively), we have $\lambda > \mu$. This places the model in the supercritical regime, where a finite fraction of letters are never answered. Numerical simulations (see supplementary information) indicate that in this supercritical regime the waiting-time distribution of the

responded letters also follows a power law with exponent $\alpha = 3/2$, which is different from the $\alpha = 1$ obtained for e-mail communications⁵. Therefore, although the response times in e-mail and mail communications follow the same scaling law, they belong to different universality classes.

The correspondence patterns of Einstein and Darwin are examples of well mapped patterns of human interaction, but are also of historical interest. Their timely responses to most letters show that they were both aware of the importance of this intellectual intercourse. Occasional delays were not always without consequence. For example, on 14 October 1921 Einstein returned to a correspondence with Theodor Kaluza that he had left off two years earlier, when he discouraged Kaluza from publishing one of his papers: having second thoughts, he recommended that the paper be submitted. Encouraged by this, Kaluza published his famous paper on five-dimensional unified field theory⁶, a key component of today's string theory. Would it have changed the course of science if Einstein had not wavered for two years? We shall never know. But our results indicate that Darwin's and Einstein's late responses or resumed correspondences are not singularities or exceptions: they are part of a universal scaling law⁷, representing a fundamental pattern of human dynamics that the famous are no better at escaping than the less distinguished.

João Gama Oliveira*†, Albert-László Barabási*‡

*Center for Complex Network Research and Department of Physics, University of Notre Dame, Indiana 46556, USA
e-mail: alb@nd.edu

†Departamento de Física, Universidade de Aveiro, 3810-193 Aveiro, Portugal

‡Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Harvard University, Boston, Massachusetts 02115, USA

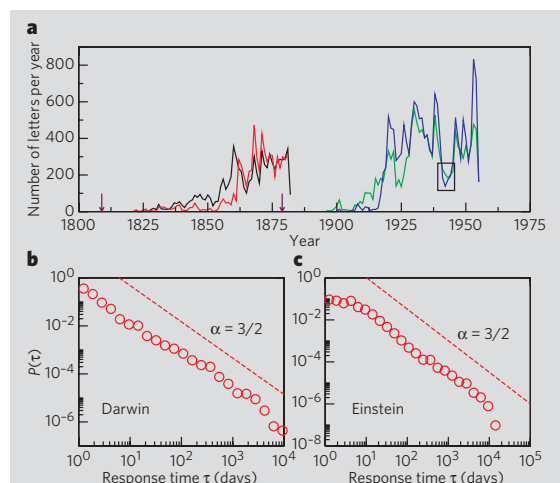


Figure 1 | The correspondence patterns of Darwin and Einstein. **a**, Historical record of the number of letters sent (Darwin, black; Einstein, green) and received (Darwin, red; Einstein, blue) each year by the two scientists^{1,2}. An anomalous drop in Einstein's correspondence marks the Second World War period (1939–45, boxed). Arrows, birth dates of Darwin (left) and Einstein (right). **b**, **c**, Distribution of response times to letters by Darwin and Einstein, respectively. Note that both distributions are well approximated with a power-law tail that has an exponent $\alpha = 3/2$, the best fit over the whole data for Darwin giving $\alpha = 1.45 \pm 0.1$ and for Einstein $\alpha = 1.47 \pm 0.1$.

1. *The Correspondence of Charles Darwin* Vols 1–14 (Cambridge Univ. Press, Cambridge, 1984–2004).
2. *The Collected Papers of Albert Einstein* Vols 1,5,8,9 (Princeton Univ. Press, New Jersey, 1993–2004).
3. Cobham, A. J. *Op. Res. Soc. Am.* **2**, 70–76 (1954).
4. Abate, J. & Whitt, W. *Queue. Syst.* **25**, 173–233 (1997).
5. Barabási, A.-L. *Nature* **435**, 207–211 (2005).
6. Kaluza, T. *Sber. Preuss. Akad. Wiss.* **54**, 966–972 (1921).
7. Bunde, A., Eichner, J. F., Havlin, S. & Kantelhardt, J. W. *Physica A* **342**, 308–314 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/4371251a

PROTEIN GLYCOSYLATION

Chaperone mutation in Tn syndrome

Tn syndrome is a rare autoimmune disease in which subpopulations of blood cells in all lineages carry an incompletely glycosylated membrane glycoprotein, known as the Tn antigen. This truncated antigen has the sugar *N*-acetyl-galactosamine α -linked to either a serine or threonine amino-acid residue^{1,2}, whereas the correct T antigen has an additional terminal galactose; the defect may be due to a malfunction of the glycosylating enzyme T-synthase^{3,4}. Here we show that Tn syndrome is associated with a somatic mutation in *Cosmc*, a gene on the X chromosome that encodes a molecular 'chaperone' that is required for the proper folding and hence full activity of T-synthase⁵. The production of the autoimmune Tn antigen by a glycosyltransferase enzyme rendered defective by a disabled chaperone may have implications for other Tn-related disorders such as IgA nephropathy, a condition that can result in renal failure.

We used whole blood from two male donors with Tn syndrome (C.C. and C.L.) and from 25 healthy donors (male and female) with a total of 33 *Cosmc* alleles between them. T-synthase activity in whole-blood cell extracts from C.C. and C.L. was significantly lower (decreased by more than 60%) than that in control samples. Tn antigens, and Tn antigens carrying additional sialic acid sugar residues, were present on erythrocytes and leukocytes from C.C. and C.L., but not on blood cells from healthy donors.

The T-synthase gene (*T-syn*; chromosome

position, 7p14–p13) contains three exons³, whereas *Cosmc* has a single exon of 954 base pairs (chromosome position, Xq23)⁵. To determine whether the defective T-synthase activity in C.C. and C.L. might be correlated with mutations in these genes, we sequenced *Cosmc* and *T-syn* from whole blood cells. *T-syn* sequences were normal for all donors, but *Cosmc* sequences from C.C. and C.L. were mosaic, containing both normal and mutated sequences. *Cosmc* from C.C. has a substitution at nucleotide 202 that gives a premature stop codon instead of an arginine residue at position 68, and a polymorphism at nucleotide 393 that causes a conservative change from aspartate to glutamate at position 131; *Cosmc* from C.L. is mutated at nucleotide 454 to give lysine instead of glutamate at position 152 (Fig. 1a, and see supplementary information).

We found that the mutation at nucleotide 202 occurred in 6 of 14 *Cosmc* clones from C.C. and that the change at nucleotide 393 was present in all 14 of his clones; C.L.'s nucleotide 454 mutation occurred in 6 of 8 clones. As *Cosmc* is X-linked and the two donors are male, these *Cosmc* sequences must be mutated in only a subset of blood cells in both. Normal *Cosmc* sequences from the 25 healthy donors, representing 33 alleles, were identical⁵. The mutation in *Cosmc* found in the two donors with Tn syndrome is statistically significant ($P < 0.01$ in Fisher's exact test).

T-synthase activity relies on coexpression with *Cosmc*⁵, so to test the effect of the *Cosmc*

mutations on the chaperone's function, we expressed recombinant *Cosmc* (wild type and mutants) together with *T-syn* in the insect cell line known as Hi-5 (Fig. 1b). As expected, co-expression of *T-syn* with *Cosmc* from C.C. that had the conservative amino-acid substitution (thymine-to-adenine polymorphism at nucleotide 393) gave normal T-synthase activity. However, coexpression with C.C.'s truncated mutant *Cosmc* (cytosine changed to thymine at nucleotide 202) gave less than 10% of the T-synthase activity associated with wild-type *Cosmc*, and no activity was detectable with C.L.'s *Cosmc* mutant. Expression of recombinant *T-syn* in Hi-5 cells was equivalent in all cases.

These results indicate that the specific mutations in *Cosmc* from patients with Tn syndrome cause it to lose its chaperone function. We confirmed by western-blot analysis that *Cosmc* protein was normally expressed from complementary DNA encoding wild-type *Cosmc* or C.C.'s polymorphic *Cosmc*. By contrast, C.C.'s truncated 68-amino-acid *Cosmc* was not detected, although C.L.'s mutant *Cosmc*, which had no chaperone activity, was detected and was normal in size.

It has been suggested that Tn syndrome is clonal and somatic^{6–8}. Our findings indicate that a somatic mutation in *Cosmc* in a subpopulation of multipotential haematopoietic stem cells in patients with Tn syndrome inhibits its chaperone activity and leads to inactivation of T-synthase and the expression of the autoimmune Tn antigen on blood cells of all lineages. This discovery may provide insight into the molecular basis for other Tn-related disorders, such as IgA nephropathy⁹ and Henoch–Schönlein purpura⁹, in which somatic mutations in *Cosmc* in haematopoietic precursors could contribute to disease aetiology.

Tongzhong Ju, Richard D. Cummings

Department of Biochemistry and Molecular Biology, and Oklahoma Center for Medical Glycobiology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104, USA

e-mail: richard-cummings@ouhsc.edu

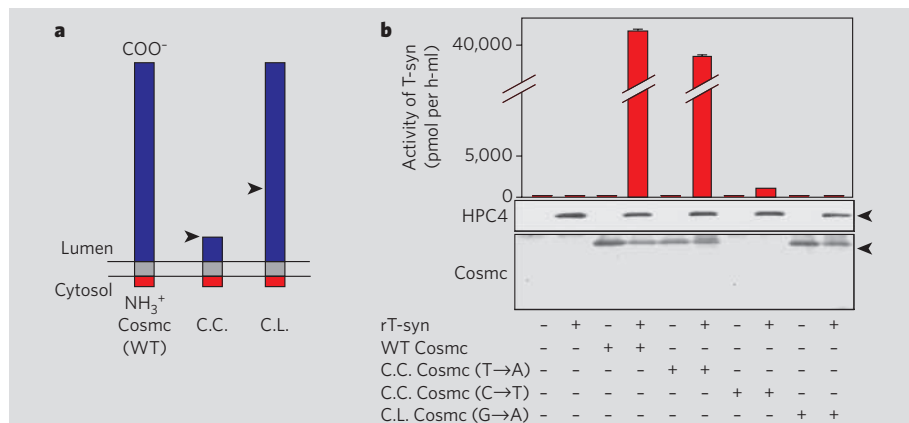


Figure 1 | Effect of *Cosmc* mutations on functional activity in patients with Tn syndrome. **a**, Length comparison of newly synthesized wild-type *Cosmc*, which is a protein of the endoplasmic reticulum that acts as a molecular chaperone for T-synthase (T-syn), with the mutated forms in patients C.C. and C.L. (arrows indicate mutation sites). **b**, Effect of coexpression of wild-type (WT) or mutant forms (C.C., + 393T→A or + 202C→T; C.L., + 454G→A, where notation numbering indicates the nucleotide mutation site and standard letter notation is used for the bases) of *Cosmc* on the activity of T-syn. Plasmids encoding these *Cosmc* variants were constructed and baculoviruses prepared in Sf-9 cells⁵ (for methods, see supplementary information). Insect Hi-5 cells were infected with baculoviruses encoding human soluble HPC4-tagged recombinant T-syn and *Cosmc*, as indicated. Top, T-syn activity in cell medium was measured in triplicate ($\pm$ s.e.m.); bottom, western blots of protein in cell medium using mouse anti-HPC4 monoclonal antibody (IgG1) to detect HPC4-tagged recombinant T-syn, and blots of protein in cell extracts using chicken anti-human *Cosmc* polyclonal antibody (IgY) to detect *Cosmc*. Migration positions of T-syn (M_r about 40K) and *Cosmc* (M_r about 37K) are indicated by arrowheads.

- Berger, E. G. *Biochim. Biophys. Acta* **1455**, 255–268 (1999).
- Cartron, J. P. & Nudelman, A. T. *Nature* **282**, 621–623 (1979).
- Ju, T., Brewer, K., D'Souza, A., Cummings, R. D. & Canfield, W. M. *J. Biol. Chem.* **277**, 178–186 (2002).
- Xia, L. et al. *J. Cell Biol.* **164**, 451–459 (2004).
- Ju, T. & Cummings, R. D. *Proc. Natl Acad. Sci. USA* **99**, 16613–16618 (2002).
- Cartron, J. P., Cartron, J., Andreu, G., Salmon, C. & Bird, G. W. *Lancet* **1**, 856–857 (1978).
- Felner, K. M., Dinter, A., Cartron, J. P. & Berger, E. G. *Biochim. Biophys. Acta* **1406**, 115–125 (1998).
- Moreau, R., Dausset, J., Bernard, J. & Moullec, J. *Bull. Mem. Soc. Med. Hop. Paris* **73**, 569–587 (1957).
- Julian, B. A. & Novak, J. *Curr. Opin. Nephrol. Hypertens.* **13**, 171–179 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing interests statement: declared none. doi:10.1038/4371252a

BRIEF COMMUNICATIONS ARISING online
 ▶ www.nature.com/bca see Nature contents.

SPORTING CONTESTS

Seeing red? Putting sportswear in context

Arising from: R. A. Hill & R. A. Barton *Nature* 435, 293 (2005)

The shirt colour worn by sportsmen can affect the behaviour of the competitors^{1,2}, but Hill and Barton³ show that it may also influence the outcome of contests. By analysing the results of men's combat sports from the Athens 2004 Olympics, they found that more matches were won by fighters wearing red outfits than by those wearing blue; they suggest that red might confer success because it is a sign of dominance in many animal species and could signal aggression in human contests. Here we use another data set from the 2004 Olympics to show that similar winning biases occur in contests in which neither contestant wears red, indicating that a different mechanism may be responsible for these effects.

If, as Hill and Barton claim, there is something special about the colour red, then contests using other colour pairings should not be biased. We tested whether this could be the case by analysing data (www.athens2004.com) from contests in which red was not used for competitors' outfits: in judo matches, one player wears blue and the other wears white. We followed the methodology originally used by Hill and Barton³ for boxing, tae kwon do, and Greco-Roman and freestyle wrestling.

After ensuring that outfit (*judogi*) colour in judo contests was allocated at random (www.ijf.org), we found a significant winning bias for players wearing blue compared with those wearing white ($\chi^2 = 7.34$, d.f. = 1, $P < 0.01$), and a similar effect of contest symmetry on winning bias to that reported by Hill and Barton (Fig. 1). We also found the winning bias for players wearing blue when considering only contests in the first round of competition ($\chi^2 = 4.85$, d.f. = 1, $P < 0.05$). This result excludes the possibility that the observed bias might arise through skilled contestants being placed, by chance, in draw positions where they wear blue more often as they progress through the competition.

Our results indicate that there is nothing

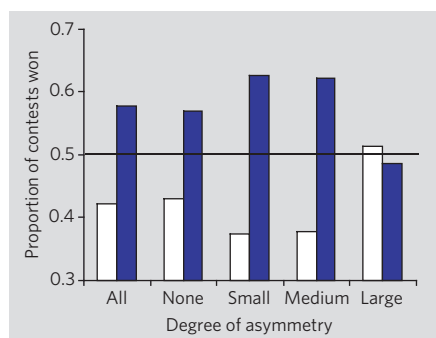


Figure 1 | Influence of *judogi* colour on the outcome of judo matches in the Athens 2004 Olympics. The black line at 0.5 indicates the expected proportion of wins by blue or white under the null hypothesis that colour has no effect on contest outcome. Blue bars, proportion won by players in blue; white bars, proportion won by players in white. There were significant differences between the number of blue and white wins for all contests combined ($\chi^2 = 7.34$, d.f. = 1, $P = 0.007$). This difference is most evident at low degrees of asymmetry in relative ability of the two competitors in each bout (subdivided using quartiles of points difference, after Hill and Barton³). Differences between the number of white and blue wins were as follows: no asymmetry ($\chi^2 = 1.53$, d.f. = 1, $P = 0.22$), small asymmetry ($\chi^2 = 6.31$, d.f. = 1, $P = 0.012$), medium asymmetry ($\chi^2 = 3.19$, d.f. = 1, $P = 0.07$) and large asymmetry ($\chi^2 = 0.57$, d.f. = 1, $P = 0.81$). Number of contests recorded are 301, 79, 99, 53 and 70, respectively.

inherently special about red in terms of colour-associated winning biases.

We can think of no plausible evolutionary explanation based on animal behaviour or evolutionary psychology that might account for a winning bias for blue contestants. We propose instead that outfit colour affects opponent visibility, which is crucial for avoidance and interception, and for anticipating behaviour. Visual abilities that could influence sporting performance include being able to

follow rapidly moving objects and perform fast visual searches⁴⁻⁶. And the hue, saturation, brightness and contrast of an object (or opponent) could enable it to be picked out against its background^{7,8}. These factors are critical for combat sports and for detecting teammates on the field of play (<http://www.liv.ac.uk/research/intelligence/issue1/manunit.html>; 1999).

In judo, the white *judogi* is likely to be perceived as brighter than the blue and may have higher contrast against the background. Men wearing blue may therefore have a visual advantage in being able to anticipate their (white) opponents' moves. We do not know the reflectance spectra, lighting arrangements or other visual factors that might have affected the visual salience of the red and blue outfits worn in the sports studied by Hill and Barton. Although our hypothesis is untested, visibility differences could also explain the biases they found. The visual attributes of sporting wear should therefore be considered in this wider perceptual context.

Candy Rowe*, Julie M. Harris†, S. Craig Roberts‡

*School of Biology & Psychology, University of Newcastle, Newcastle upon Tyne NE2 4HH, UK
e-mail: candy.rowe@ncl.ac.uk

†School of Psychology, University of St Andrews, St Mary's College, St Andrews KY16 9JP, UK

‡Evolutionary Psychology & Behavioural Ecology Group, School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK

1. Frank, M. G. & Gilovich, T. J. *Pers. Soc. Psychol.* **54**, 74–85 (1988).
2. Mills, B. D. & French, L. M. J. *Hum. Mov. Stud.* **31**, 47–60 (1996).
3. Hill, R. A. & Barton, R. A. *Nature* **435**, 293 (2005).
4. Williams, A. M. J. *Sport Sci.* **18**, 737–750 (2000).
5. Farrow, D. & Southgate D. *Clin. Exp. Optom.* **83**, 226–231 (2000).
6. Williams, M. *Psychologist* **15**, 416–417 (2002).
7. McKeefry, D. J. et al. *Invest. Ophthalm. Vis. Sci.* **44**, 2267–2276 (2003).
8. Regan, B. C. et al. *Phil. Trans. R. Soc. Lond. B* **356**, 229–283 (2001).

doi:10.1038/nature04306

SPORTING CONTESTS

Hill & Barton reply

Replying to: C. Rowe, J. M. Harris and S. C. Roberts *Nature* 437, doi:10.1038/nature04306 (2005)

Rowe *et al.* corroborate our finding that the colour of clothing influences the outcome of sporting contests¹, but they offer a different mechanism to explain the effect. We found that in four combat sports wearing red was consistently associated with improved perfor-

mance relative to wearing blue², and argued that wearing red enhances performance through psychological effects on the wearer and/or on the opponent. We suggested that these psychological effects reflect the evolutionary and cultural associations of red with

dominance and aggression. Rowe *et al.* find that, in a fifth combat sport, wearers of blue outperformed wearers of white. They attribute their and our results to perceptual rather than to psychological effects, arguing that visibility of the opponent is the critical factor.

In our view, this visibility explanation is unlikely in a situation where contestants fight at close quarters in brightly lit arenas, as in these combat sports. Crucially, in the combat sports we analysed, the hypothesis of Rowe *et al.* requires that blue-wearing opponents be more visible than their red-wearing opponents,

but, insofar as photographs provide an indication of this, the opposite seems to be true (see <http://www.athens2004.com/en/BoxingImageGallery/imagegallery>). Furthermore, the literature on visibility effects cited by Rowe *et al.* refers to sports where visuospatial judgements are made over distance. We also note that visibility has not been used to explain similar effects of artificially coloured leg bands on dominance interactions in birds³.

The visibility explanation proposed by Rowe *et al.* suggests that male and female contests should be equally affected by colour, whereas our sexual-selection hypothesis would predict stronger effects for males. The latter prediction is supported by the data: in the same sports in which significant effects were found for males, no such effects are evident for females (tae kwon do and freestyle wrestling (red against blue): $\chi^2 = 0.32$, $n = 155$, d.f. = 1, $P > 0.50$; judo (blue against white): $\chi^2 = 0.30$, $n = 214$, d.f. = 1, $P > 0.50$). Pooling

the data for the three combat sports in which both men and women participated, there is a significant association between sex and proportion of contests won by the advantageous colour ($\chi^2 = 6.44$, d.f. = 1, $P = 0.011$).

We suggest that the impact of colour operates through its psychological and hormonal influences, rather than through its effects on visibility. Why then should blue provide an advantage over white? Although red reflects current dominance and testosterone levels⁴ and indicates emotional arousal⁵, other bright colours, including blue, are known to indicate long-term developmental vigour⁶ and may therefore have psychological effects relative to paler colours or white. In any case, our demonstration of enhanced performance in red relative to blue and other colours across a range of conditions suggests that, contrary to Rowe *et al.*'s claim, there is indeed something special about red.

Identification of the precise mechanisms

underlying the effects of colour will require more detailed experimental work. Meanwhile, the results of Rowe *et al.* lend further support to our claim that colour needs to be taken into account in ensuring a level playing field in sport.

Robert A. Barton, Russell A. Hill

Evolutionary Anthropology Research Group,
Department of Anthropology, Durham University,
Durham DH1 3HN, UK
e-mail: r.a.hill@durham.ac.uk

1. Rowe, C., Harris, J. M. & Roberts, S. C. *Nature* **437**, doi:10.1038/nature04306 (2005).
2. Hill, R. A. & Barton, R. A. *Nature* **435**, 293 (2005).
3. Cuthill, I. C. *et al. Proc. R. Soc. Lond. B* **264**, 1093–1099 (1997).
4. Setchell, J. M. & Wickings, E. J. *Ethology* **111**, 25–50 (2005).
5. Drummond, P. D. & Quah, S. H. *Psychophysiology* **38**, 190–196 (2001).
6. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, 1994).

doi:10.1038/nature04307



SLEEP

Cover illustration

Top right: A lamp illuminates the face of a sleeping man (Credit: Corbis). Top left: Tracings from an electroencephalogram recording eye movements during rapid eye movement sleep (Credit: M. Mahowald and C. H. Schenck). Bottom: A man asleep at this desk (Credit: B. Van der Meer/Getty).

Editor, *Nature*

Philip Campbell

Insights Publisher

Sarah Greaves

Insights Editor

Lesley Anson

Production Editor

Maria Hodges

Senior Art Editor

Martin Harrison

Art Editor

Nik Spencer

Layouts

Marie-Claire Patin

Sponsorship

Claire Hines
Claudia Banks

Production

Sue Gray

Marketing

Claire Aspinall

Editorial Assistant

Laura Shaw

The fundamental truths of sleep are not difficult to master: one sleeps when one is tired — mostly at night—and awakens the next day usually feeling rested and refreshed.

So why put together an Insight on a topic that seems so straightforward?

Although it is often true in biology that things are more complex than they seem at first glance, it is especially accurate for sleep. This became apparent about 50 years ago with the discovery of rapid eye movement (REM) sleep. This is a sleep state marked by intense brain activity, rapid bursts of eye movement and vivid dreaming. The high level of brain activity during REM sleep created a serious challenge to the prevailing dogma — that we sleep simply to provide rest — and raised a host of largely unanswered questions about the function of sleep.

Intuition also fails us when considering other aspects of sleep — namely that ‘drifting off to sleep’ is a slow process and that sleep and wake are completely separate states. On the contrary, the act of switching from being awake to sleeping can be extremely rapid, an observation that carries significant public health implications. And patients with various sleep disorders can exist in curious states that combine aspects of both sleep and wakefulness, indicating that the two are not always mutually exclusive.

That so many big questions in sleep research remain unanswered makes it a fascinating field to follow. This Insight highlights much of that excitement with a diverse collection of articles.

We are pleased to acknowledge the support of the National Institutes of Health in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

John Spiro, Senior Editor

COMMENTARY

1254 Sleep is of the brain, by the brain and for the brain

J. A. Hobson

REVIEWS

1257 Hypothalamic regulation of sleep and circadian rhythms

C. D. Saper, T. E. Scammell & J. Lu

1264 Clues to the functions of mammalian sleep

J. M. Siegel

1272 Sleep-dependent memory consolidation

R. Stickgold

1279 Insights from studying human sleep disorders

M. W. Mahowald & C. H. Schenck

PROGRESS

1286 What are the memory sources of dreaming?

T. A. Nielson & P. Stenstrom

nature
insight

Sleep is of the brain, by the brain and for the brain

J. Allan Hobson¹

Sleep is a widespread biological phenomenon, and its scientific study is proceeding at multiple levels at the same time. Marked progress is being made in answering three fundamental questions: what is sleep, what are its mechanisms and what are its functions? The most salient answers to these questions have resulted from applying new techniques from basic and applied neuroscience research. The study of sleep is also shedding light on our understanding of consciousness, which undergoes alteration in parallel with sleep-induced changes in the brain.

For decades it was assumed that brain activity was greatly reduced or absent during sleep. Subjective experience of the loss of consciousness and the lack of memory of mental activity during sleep appeared to support this conclusion. Even such great scientists as Charles Sherrington¹ and Ivan Pavlov² backed this idea.

This assumption was overturned when the regular cyclic alteration of rapid eye movement (REM) and non-REM (NREM) sleep phases was discovered in the 1950s and 60s (ref. 3). The discovery of REM sleep and its correlation with vivid hallucinatory dreaming was evidence that the brain was highly active during sleep⁴. Soon after this discovery it was noticed that sensory inputs and motor outputs were simultaneously blocked when the brain was activated during REM sleep, putting it 'off-line'.

It was a great surprise to discover that the vigorous brain activation of REM sleep occurred at regular 90-minute intervals and occupied up to 20% of sleep. This fact alone invalidated the belief that sleep was caused by and associated with a cessation of brain activity. Other facts supported the idea that the brain was continuously active during sleep. The early cerebral blood flow studies of Kety⁵ and later Sokolov showed only a 20% reduction in cerebral blood flow during sleep. Because blood flow is correlated with neuronal activity it should not have been a surprise to find that almost as many neurons increased their firing rate at sleep onset as their activity decreased⁶. Even during NREM sleep, when consciousness may be totally obliterated, the brain remains significantly active.

The descriptive study of sleep, so richly productive during the early years of the sleep laboratory era, has been complemented in the past decade by the application of brain imaging^{7–9} and quantitative electroencephalogram (EEG) mapping¹⁰. Imaging techniques showed that the regional activation of the brain is very different in the two EEG-activated states, REM sleep and waking, and both are different from the NREM phase of sleep (when the EEG shows high-voltage slow waves instead of low-voltage fast activity). Quantitative EEG studies also revealed regional differences in brain electrical activity^{10,11}.

These new data indicate that the brain is relatively quiescent during slow-wave sleep (when the EEG is dominated by sleep spindles and high-voltage slow waves). But it must be emphasized that such global deactivation is only relative. Although consciousness is dulled, the brain is still roughly 80% activated and thus capable of robust and elaborate

information processing. Thus, the EEG spindles and slow waves represent changes in the excitability of cortical and thalamic circuitry and should be regarded not simply as 'noise', which subjective experience leads us to assume, but as signals used by the brain for its own functional purposes¹².

Mechanisms and functions of sleep

All in all, these findings support two radical ideas. One is that sleep is an actively regulated process, not simply the passive result of diminished waking. The other is that sleep should be regarded as a reorganization of neuronal activity rather than a cessation of activity. With respect to the first idea, it soon became apparent that although mammalian sleep occurs during the rest phase of the circadian rhythm, it is produced by brain processes in the hypothalamus and brainstem. It is in this context that Saper's recent description of a sleep switch is best understood and appreciated (see the review in this issue by Saper, Scammell and Lu, p. 1257).

Almost all mammals that have been studied show the NREM–REM cyclic alternation, which suggests not only a shared mechanism across species but a universal functional significance that must be far richer than the mere energy saving that the subjective experience-based theories thought adequate. The discovery of the continuous and elaborately modulated nature of sleep made it imperative to seek more active functional consequences, such as the homeostatic control of energy and the reinforcement of learning, that have recently been found. This is the context for thinking about Stickgold's descriptions of procedural learning enhancement during sleep (see the review in this issue by Stickgold, p. 1272).

The resurgence of interest in sleep and learning was sparked by robust evidence for sleep's promotion of consolidation and improvement in learned motor skill performance in terrestrial mammals. Some animals can learn despite having little and/or poor sleep, but this does not mean that demonstrated sleep–learning links are irrelevant artefacts. The general principle could be that a species uses sleep for learning if it can afford to do so. The more vexing problem of understanding how sleep does (or does not) benefit narrative memory will be solved only by more assiduous and strategic study. Then, and only then, can we speak of sleep as beneficial to memory where memory is defined as the conscious recollection of learned

¹Department of Psychiatry, Harvard Medical School, 74 Fenwood Road, 401 Park Drive., 2nd Floor East, Boston, Massachusetts 02115, USA.

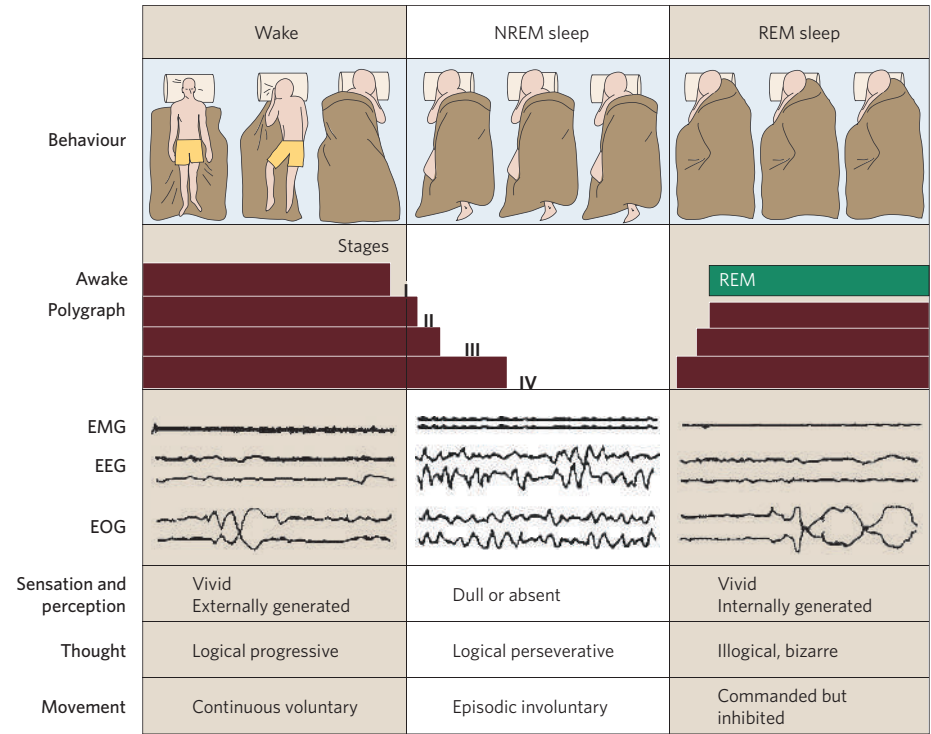


Figure 1 | Behavioural states in humans. States of waking, NREM sleep and REM sleep have behavioural, polygraphic and psychological manifestations. In the row labelled behaviour, changes in position (detectable by time-lapse photography or video) can occur during waking and in concert with phase changes of the sleep cycle. Two different mechanisms account for sleep immobility. The first is disfacilitation (during stages I–IV of NREM sleep). The second is inhibition (during REM sleep). During dreams, we imagine that we move, but we do not. Sample tracings of three variables used to distinguish the state are shown: an electromyogram (EMG), an electroencephalogram (EEG) and an electrooculogram (EOG). The EMG tracings are highest during waking, intermediate during NREM sleep and lowest during REM sleep. The EEG and EOG are both activated during waking and inactivated during NREM sleep. Each sample shown is approximately 20 seconds long. The three bottom rows describe other subjective and objective state variables. Modified from ref. 19.

material. Memory, so defined, depends upon learning but is not equivalent to it.

The variation in sleep between species and during their lifespan suggests, further, that sleep may have many functions. And those functions may not only vary but be absent in some animals. The same functions may also be achieved during periods of waking. But this flexibility is not evidence of functionlessness. In other words, to say that sleep is variable does not, of course, mean that sleep is not vital to those species that do sleep. It only means that relatively sleepless species have some other way of adapting to life's demands. Such plasticity is evidence of a pluralistic and, we assume, adaptive set of brain mechanisms and functions associated with sleep. This is the idea that inspires and underpins Siegel's phylogenetic work (see the review by Siegel, p. 1264).

The study of sleep phylogeny has a long history. It soon became clear that sleep, as we know it in higher mammals, tends to be correlated with relatively large brains and with homeothermy. The searching studies of Allison & Cichetti¹³ demonstrated that adaptations to diverse ecological niches also played a major role in determining the amount, temporal distribution and depth of sleep. The general rule was that large, carnivorous surface-dwelling animals, such as lions, slept long and deep whenever they were not foraging or mating. By contrast, smaller, herbivorous species such as rabbits tend to be nest dwellers and sleep relatively little. They have frequent awakenings and spend more time foraging and eating. They needed to be vigilant to defend themselves from predators. The commonsense conclusion from this work is that an animal sleeps if it can afford to.

Additional evidence of variability comes from the studies of human sleep that have uncovered a vast and complex set of sleep disorders. If one considers sleep to be an actively regulated process, it is understandable that some people get too little, others too much and others the wrong kind of sleep. The functional consequence of these disorders is similarly variable. Mahowald and Schenks's review (see p. 1279 in this issue) of these sleep disorders emphasizes some of the many possible dissociations of sleep and wake components that can afflict us.

What effect has the new science of sleep had on dream theory? Contrary to Freud's assertion that dreaming was stimulated by mem-

ory of that same day's experience (ref. 14), new data indicate that an equally large incorporation of recollected memory antedates its expression in dreams by as much as six days. While awaiting systematic replication, this finding should caution us about the uncritical acceptance of any theoretical formation about dreaming that is not based on evidence. A case in point is the finding that whatever the time lag to incorporation of some recent events in dreams, most dream content has no identifiable experiential antecedent¹⁵. Nielsen's studies pursue the implications of some of these findings (see the review in this issue, p. 1286).

Sleep and consciousness

Perhaps the most far-reaching of the complications of modern sleep science concerns the riddle of the basis of consciousness, a theme not addressed by any of the articles in this Insight.

A moment's reflection supports the idea that consciousness is state dependent. For centuries we made judgements about sleep and the brain that were wrong because we assumed, mistakenly, that consciousness ceased at sleep onset and resumed only when we woke. The occasional recall of dreams should have ruined that theory but great minds, including that of Sigmund Freud, incorrectly assumed that dreaming only occurred during the process of awakening.

Although it is true that consciousness is dulled during deep NREM sleep, it is qualitatively altered in parallel with the reorganization of brain activity that occurs at sleep onset when dreamlike mental activity is fleeting. This takes place during the lighter stages of NREM sleep, when dreaming may be more sustained and, most markedly, during REM sleep when dreaming assumes its most florid character. Because memory is so severely affected by sleep it has been difficult to get reliable and valid descriptions of mental activity during sleep, but it is now clear that consciousness undergoes alteration in parallel with sleep changes in the brain.

The net effect of this conclusion is to strengthen the conscious state hypothesis, which asserts that consciousness changes its intensity and character in a stereotypical way as the brain changes state during the sleep-wake cycle. One approach to the scientific study of consciousness is to simultaneously track changes in the brain and

changes in the mind and then to map back and forth between the two domains. How might brain activity mediate conscious experience? The direct study of subjective experience during sleep is an important part of the current flurry of excitement in the domain of consciousness research¹⁶. Subjective experience is so problematical that all but the bravest scientists¹⁷ have been discouraged. And yet it must be recognized that subjectivity cannot be studied at all unless pains are taken to overcome the pitfalls of using first-person data. When large samples are taken, and the level of analysis is coarse-grained and its focus formal, the subjective data correlate with brain data at high levels of statistical and functional significance. Two examples help to make this point.

Normal subjects were asked to report their mental experience during periods of active wake, quiet wake, sleep onset, NREM sleep and REM sleep. The reports were scored for descriptions of hallucinatory perception and of thinking, and a reciprocal relation to brain state was observed. Hallucinatory mental content is lowest during active waking and highest during REM sleep. The incidence of thinking is reciprocally highest during quiet waking and lowest during REM sleep¹⁸.

The implication of these findings is that the sleeping brain can either generate its own perceptions or it can think about them. It cannot do both at the same time. Dreaming is therefore as hallucinatory and thoughtless (or delusional) as so-called mental illness.

In the second study we tested this hypothesis. When psychotic schizophrenic patients were given the thematic apperception test (TAT), in which verbal descriptions of simple but ambiguous pictures are recorded and scored, when they were awake and asked to report their dreams, they had equally high scores on a bizarreness scale (designed to pick up cognitive discontinuity and incongruity) for both. Age- and sex-matched normal control subjects have the same amount of dream bizarreness as the patients but are much less bizarre in their wake-state projective test responses (S. Scarone, M. L. Manzone, O. Gambini and J. A. Hobson, unpublished data).

These findings support the hypothesis that REM sleep is a physiological brain state that produces a distinctive and psychosis-like mental content, whereas during normal waking such properties are suppressed. Put another way, when awake the brain is normally free of the formal aspects of dream activity. Conversely, normal dreaming is justifiably considered to be an entirely normal model of highly abnormal conditions of the human brain and mind. It is now clear that the

kind of consciousness that a person experiences is a function of the state of the brain.

Sleep and dream research is a rare convergence point for the biological and psychological sciences. Ongoing work in this area promises to bridge areas of research from molecular and cell biology, through neuronal populations, to behavioural and conscious states. It may even prove helpful in solving the mind-body problem. ■

1. Sherrington, C. *Man on his Nature* (Doubleday, Garden City, New York, 1995).
2. Pavlov, I. I. *Conditioned Reflexes. An Investigation of the Physiological Activity of the Cerebral Cortex* (Dover, New York, 1960).
3. Aserinsky, E. & Kleitman, N. Regularly occurring periods of eye motility and concomitant phenomena during sleep. *Science* **118**, 273–274 (1953).
4. Dement, W. C. & Kleitman, N. The relation of eye movements, body motility, and external stimuli to dream content. *J. Exp. Psychol.* **55**, 543–553 (1957).
5. Kety, S. S., Landau, W. M., Freygang, W. H., Rowland, L. P. & Sokoloff, L. The local circulation of the living brain; values in the unanesthetized and anesthetized cat. *Trans. Am. Neurol. Assoc.* **80**, 125–129 (1955).
6. Hobson, J. A. *Dreaming: An Introduction to the Science of Sleep* (Oxford Univ. Press, New York, 2002).
7. Maquet, P. et al. Functional neuroanatomy of human rapid-eye movement sleep and dreaming. *Nature* **383**, 163–166 (1996).
8. Nofzinger, E. A., Mintun, M. A., Wiseman, M. B., Kupfer, D. J. & Moore, R. Y. Forebrain activation in REM sleep: An FDG PET study. *Brain Res.* **770**, 192–201 (1997).
9. Braun, A. R. et al. Regional cerebral blood flow throughout the sleep-wake cycle — an (H₂O)-O-15 PET study. *Brain* **120**, 1173–1197 (1997).
10. Achermann, P. & Borbély, A. A. Low frequency (<1 Hz) oscillations in the human sleep electroencephalogram. *Neuroscience* **81**, 213–222 (1997).
11. Huber, R., Ghilardi, M. F., Massimini, M. & Tononi, G. Local sleep and learning. *Nature* **430**, 78–81 (2004).
12. Steriade, M., Timofeev, I. & Grenier, F. Intracellular activity of various neocortical cell-classes during the natural wake-sleep cycle. *Soc. Neurosci. Abstr.* **25**, 1661 (1999).
13. Allison, T. & Cicchetti, D. V. Sleep in mammals. Ecological and constitutional correlates. *Science* **194**, 732–734 (1976).
14. Freud, S. *An Outline of Psychoanalysis* (Hogarth Press, London, 1949).
15. Fosse, M., Fosse, R., Hobson, J. A. & Stickgold, R. Dreaming and episodic memory: a functional dissociation? *J. Cogn. Neurosci.* **15**, 1–9 (2003).
16. Libet, B., Gleason, C. A., Wright, E. W. and Pearl, D. K. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* **106**, 623–642 (1983).
17. Metzinger, T. *On Being No One* (MIT, Cambridge MA, 2003).
18. Fosse, R., Stickgold, R. & Hobson, J. A. Brain-mind states: reciprocal variation in thoughts and hallucinations. *Psychol. Sci.* **12**, 30–36 (2001).
19. Hobson, J. A. & Steriade, M. In *Handbook of Physiology: The Nervous System Vol. 4* (eds Mountcastle, V. and Bloom, F. E.) 701–823. (Am. Physiol. Soc., Bethesda, 1986).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to J.A.H. (allan_hobson@hms.harvard.edu).

Hypothalamic regulation of sleep and circadian rhythms

Clifford B. Saper¹, Thomas E. Scammell¹ & Jun Lu¹

A series of findings over the past decade has begun to identify the brain circuitry and neurotransmitters that regulate our daily cycles of sleep and wakefulness. The latter depends on a network of cell groups that activate the thalamus and the cerebral cortex. A key switch in the hypothalamus shuts off this arousal system during sleep. Other hypothalamic neurons stabilize the switch, and their absence results in inappropriate switching of behavioural states, such as occurs in narcolepsy. These findings explain how various drugs affect sleep and wakefulness, and provide the basis for a wide range of environmental influences to shape wake–sleep cycles into the optimal pattern for survival.

In 1916, Baron Constantin von Economo, a Viennese neurologist, began to see patients with a new type of encephalitis that specifically attacked regions of the brain that regulate sleep and wakefulness¹. This disorder, which was eventually called encephalitis lethargica or von Economo's sleeping sickness, swept through Europe and North America during the second decade of the twentieth century; by the end of the following decade it had apparently disappeared, as only sporadic and unconvincing reports have appeared since. Although the virus that caused it was never identified, von Economo was able to identify the areas of the brain in which lesions caused specific alterations of wake–sleep regulation (Fig. 1). Only during the past decade has the accuracy of his observations come to be appreciated, as key components of the wake–sleep-regulatory system were found to reside at the sites that von Economo first identified.

In this review, we cover recent advances in understanding the brain circuitry that regulates sleep and produces wakefulness, including cell groups in the brainstem, hypothalamus and basal forebrain (BF) that are crucial for arousing the cerebral cortex and thalamus. These neurons are inhibited during sleep by a system of γ -aminobutyric acid (GABA)-containing neurons, in which the ventrolateral preoptic nucleus (VLPO) seems to have a key role. Mutual inhibition between the arousal- and sleep-producing circuitry results in switching properties that define discrete wake and sleep states, with sharp transitions between them. We examine how this switch is stabilized by orexin neurons in the lateral hypothalamus (LHA), and why loss of these neurons results in narcolepsy. We also explore how basic drives affect this system, including the homeostatic need to sleep that accumulates as a function of prolonged wakefulness, the circadian drive that moulds the daily cycles of sleep–wakefulness, and allostatic influences that adapt sleep and wakefulness to external behavioural events. Interactions with this circuitry might explain the effects of a range of drugs on sleep and wakefulness.

The ascending arousal system promotes wakefulness

The majority of patients with von Economo's encephalitis lethargica slept excessively. Many slept for 20 or more hours per day, arising only briefly to eat and drink. Their cognitive function was reasonably intact, but they would soon return to sleep — a cycle that lasted for

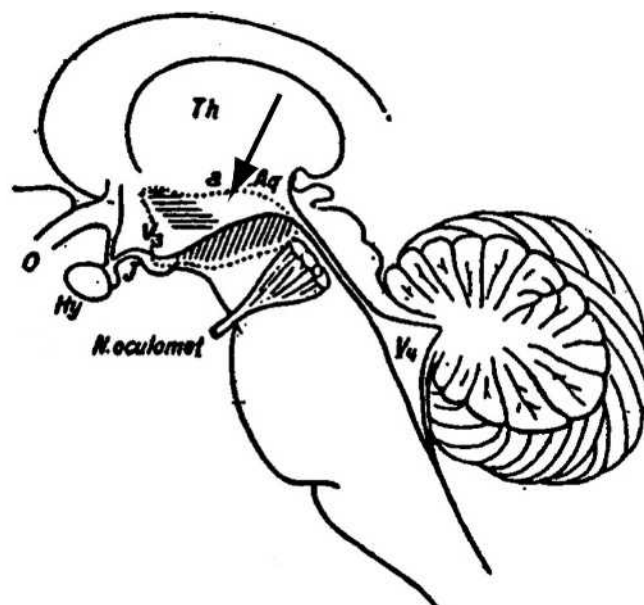


Figure 1 | A drawing of the human brainstem, taken from von Economo's original work. It illustrates the site of the lesion (diagonal hatching) at the junction of the brainstem and forebrain that caused prolonged sleepiness, and the site of the lesion (horizontal hatching) in the anterior hypothalamus that caused prolonged insomnia. The arrow points to a region between the two, including the posterior lateral hypothalamus. Von Economo suggested that narcolepsy was caused by lesions at this site.

many weeks before recovery¹. Von Economo found that these patients invariably had lesions at the junction of the midbrain and the diencephalon. He therefore proposed that there was an ascending arousal system originating in the brainstem that kept the forebrain awake (Fig. 1). During the years after the Second World War, investigators, such as Moruzzi and Magoun and their many colleagues, showed that this influence was mediated by an ascending arousal pathway that begins in the rostral pons and runs through the midbrain reticular for-

¹Department of Neurology and Program in Neuroscience, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, Massachusetts, 02215, USA.

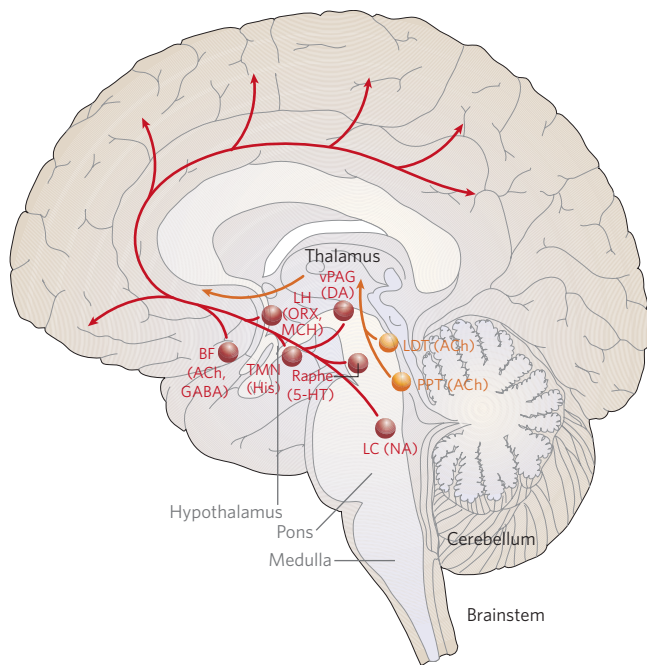


Figure 2 | A schematic drawing showing some key components of the ascending arousal system. A major input to the relay and reticular nuclei of the thalamus (yellow pathway) originates from cholinergic (ACh) cell groups in the upper pons, the pedunculopontine (PPT) and laterodorsal tegmental nuclei (LDT). These inputs facilitate thalamocortical transmission. A second pathway (red) activates the cerebral cortex to facilitate the processing of inputs from the thalamus. This arises from neurons in the monoaminergic cell groups, including the tuberomammillary nucleus (TMN) containing histamine (His), the A10 cell group containing dopamine (DA), the dorsal and median raphe nuclei containing serotonin (5-HT), and the locus coeruleus (LC) containing noradrenaline (NA). This pathway also receives contributions from peptidergic neurons in the lateral hypothalamus (LHA) containing orexin (ORX) or melanin-concentrating hormone (MCH), and from basal forebrain (BF) neurons that contain γ -aminobutyric acid (GABA) or ACh. Note that all of these ascending pathways traverse the region at the junction of the brainstem and forebrain where von Economo noted that lesions caused profound sleepiness.

mation²; hence, the concept of the 'ascending reticular activating system' achieved wide currency^{2,3}.

Studies in the 1970s and 1980s clarified the nature of this pathway (Fig. 2). A key finding was that the ascending arousal system largely originates from a series of well-defined cell groups with identified neurotransmitters⁴. This pathway has two major branches. The first branch is an ascending pathway to the thalamus that activates the thalamic relay neurons that are crucial for transmission of information to the cerebral cortex. The major source of upper brainstem input to the thalamic-relay nuclei, as well as to the reticular nucleus of the thalamus, is a pair of acetylcholine-producing cell groups: the pedunculopontine and laterodorsal tegmental nuclei (PPT/LDT)⁵. The neurons in the PPT/LDT fire most rapidly during wakefulness and rapid eye movement (REM) sleep, which is the stage accompanied by cortical activation, loss of muscle tone in the body and active dreams⁶. These cells are much less active during non-REM (NREM) sleep, when cortical activity is slow. Their input to the reticular nucleus is crucial, as it sits between the thalamic-relay nuclei and the cerebral cortex, acting as a gating mechanism that can block transmission between the thalamus and cerebral cortex, which is important for wakefulness⁷. Other inputs to the thalamic midline and intralaminar nuclei originate more broadly in the upper brainstem, including the reticular formation, the PPT/LDT, the monoaminergic systems (see below) and the parabrachial nucleus⁸. The intralaminar and midline nuclei are also believed to have a role in cortical arousal.

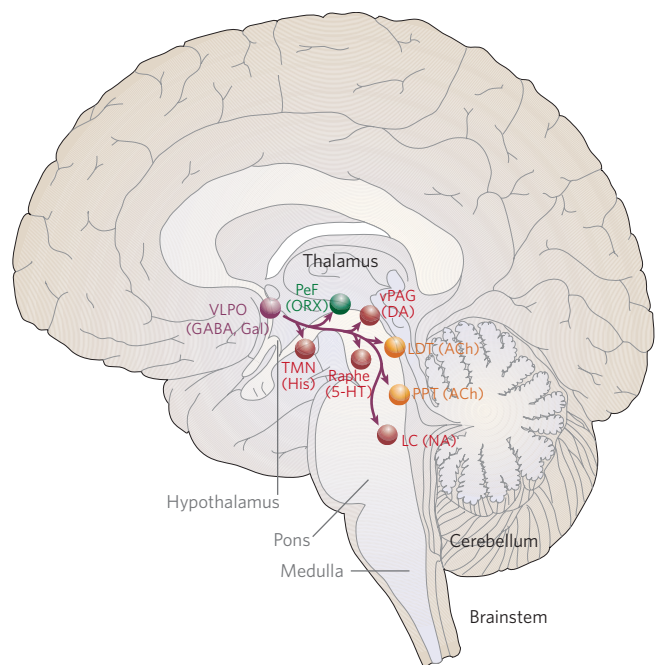


Figure 3 | A schematic drawing to show the key projections of the ventrolateral preoptic nucleus (VLPO) to the main components of the ascending arousal system. It includes the monoaminergic cell groups (red) such as the tuberomammillary nucleus (TMN), the A10 cell group, the raphe cell groups and the locus coeruleus (LC). It also innervates neurons in the lateral hypothalamus (LHA; green), including the perifornical (PeF) orexin (ORX) neurons, and interneurons in the cholinergic (ACh) cell groups (yellow), the pedunculopontine (PPT) and laterodorsal tegmental nuclei (LDT). Note that the VLPO neurons lie within the region outlined by von Economo for the anterior hypothalamic lesion that caused insomnia. 5-HT, serotonin; GABA, γ -aminobutyric acid; gal, galanin; NA, noradrenaline; His, histamine.

The second branch of the ascending arousal system bypasses the thalamus, instead activating neurons in the lateral hypothalamic area and BF, and throughout the cerebral cortex^{4,9,10}. This pathway originates from monoaminergic neurons in the upper brainstem and caudal hypothalamus, including the noradrenergic locus coeruleus (LC), serotonergic dorsal (DR) and median raphe nuclei, dopaminergic ventral periaqueductal grey matter and histaminergic tuberomammillary neurons. The input to the cerebral cortex is augmented by lateral hypothalamic peptidergic neurons (containing melanin-concentrating hormone (MCH) or orexin/hypocretin), and BF neurons (containing acetylcholine or GABA). Lesions along this pathway, particularly in the LHA and rostral midbrain, produce the most profound and long-lasting forms of sleepiness or even coma^{11,12}. Neurons in each of the monoaminergic nuclei that contribute to this pathway have the property of firing fastest during wakefulness, slowing down during NREM sleep and stopping altogether during REM sleep^{13–15}. Orexin neurons in the LHA are, similarly, most active during wakefulness^{16–18}, whereas MCH neurons are active during REM sleep¹⁹. Many BF neurons, including most cholinergic neurons, are active during both wake and REM sleep²⁰.

Given the anatomy and functions of these systems, it is not surprising that von Economo found that lesions at the junction of the midbrain and forebrain, which block both ascending pathways, produce profound and long-lasting impairment of arousal.

The VLPO promotes sleep

Von Economo's second major observation was an opposite response in a small percentage of the victims of encephalitis lethargica¹. Rather than being sleepy, they became insomniac and slept for only a few hours each day. Typically, they were extremely tired, but found it diffi-

cult to fall asleep, could sleep for only a short time, and then awoke and were unable to fall back asleep. These patients had lesions involving the basal ganglia and adjacent anterior hypothalamus. Later experiments in animals identified a hypothalamic site involving the lateral preoptic area where lesions caused similar insomnia^{21,22}.

During the 1980s and 1990s, investigators began to examine the inputs to the monoaminergic cell groups that might be responsible for their remarkable, stereotyped and coordinated changes in firing patterns associated with sleep. One such cell group, the VLPO, was found to send outputs to all of the major cell groups in the hypothalamus and brainstem that participate in arousal²³ (Fig. 3). The VLPO neurons are primarily active during sleep, and contain the inhibitory neurotransmitters, galanin and GABA^{24–26}. These neurons form a dense cluster, as well as a more diffuse extended part of the nucleus^{24,25}.

These observations suggested that damage to the VLPO might have caused the insomnia in von Economo's patients. Experiments showed that cell-specific lesions of the VLPO in animals reduced both NREM and REM sleep by more than 50% (refs 27, 28). Lesions of the VLPO cluster primarily reduced NREM sleep, whereas lesions of the extended VLPO mainly disrupted REM sleep. The extended VLPO neurons provide the main output from the VLPO to the LC and DR, which are thought to be important in gating REM sleep^{27,29}. By contrast, the VLPO cluster more heavily innervates the histaminergic neurons, which are closely linked to transitions between arousal and NREM sleep^{27,30,31}.

The VLPO also receives afferents from each of the major monoaminergic systems³². Both noradrenaline and serotonin inhibit VLPO neurons³³. The latter do not have histamine receptors, but the tuberomammillary neurons also contain GABA³⁴, which is inhibitory to VLPO neurons³⁵, as well as several other potentially inhibitory peptides, such as galanin and endomorphin^{36,37}. Therefore, the VLPO can be inhibited by the very arousal systems that it inhibits during sleep (Fig. 4).

The flip-flop switch

A circuit containing mutually inhibitory elements sets up a self-reinforcing loop, where activity in one of the competing sides shuts down inhibitory inputs from the other side, and therefore disinhibits its own action. Such a circuit is called a 'flip-flop switch' by electrical engineers, who design them to produce discrete states with sharp transitions⁴. Flip-flop circuits tend to avoid transitional states, because when either side begins to overcome the other, the switch 'flips' into the alternative state. This flip-flop circuit model might explain why wake-sleep transitions are often relatively abrupt (one 'falls' asleep and suddenly awakes), and both humans and animals spend only a small part of each day (typically < 1–2%) in transitional states (Fig. 4). There are obvious adaptive advantages to a wake-sleep system that makes rapid and complete transitions, as it would be dangerous for animals to have impaired alertness while engaging in waking behaviours, and inefficient for them to spend their sleep periods half-awake.

However, flip-flop switches can also make unwanted transitions with little warning. When a small perturbation gives one side a sudden advantage, it can turn off the alternative state relatively abruptly (such as falling asleep during a momentary lapse of attention while driving). Interestingly, mathematical models show that when either side of a flip-flop neural circuit is weakened, homeostatic forces cause the switch to ride closer to its transition point during both states³⁸. As a result, there is an increase in transitions, both during the wake and the sleep periods, regardless of which side is weakened. This is certainly seen in animals with VLPO lesions, which fall asleep about twice as often as normal animals, wake up much more often during their sleep cycle and, on the whole, only sleep for about one-quarter as long per bout²⁸ — in other words, they wake up and are unable to fall back asleep during the sleep cycle, but also are chronically tired, falling asleep briefly and fitfully during the wake cycle, much like von Economo's patients. Interestingly, elderly people, who have similar problems although perhaps on a lesser scale³⁹, have been found to have

loss of neurons in the human equivalent of the VLPO²⁴ with ageing^{40–42}.

Orexin/hypocretin neurons and state stability

In the years that followed the epidemic of encephalitis lethargica, several well-known clinicians, such as Wilson in London and Spiller in Philadelphia, recorded a rash of cases of narcolepsy, a mysterious illness in which the patients had attacks of irresistible sleepiness, as well as episodes of cataplexy during which emotional situations would cause them to lose muscle tone and collapse to the floor⁴³. Wilson believed that at least some of the cases were post-encephalitic, and von Economo proposed that the responsible pathology involved the posterior hypothalamus¹.

In 1998, two groups of investigators simultaneously discovered a pair of closely related neuropeptides, named orexins by one group and hypocretins by the other^{44–46}. These peptides are produced exclusively by a cluster of neurons in the posterior half of the LHA. One year later, two groups of investigators again simultaneously found that a lack of orexins or their type 2 receptor can cause symptoms of narcolepsy in experimental animals^{46,47}. The following year, it was reported that humans who have narcolepsy with cataplexy have few orexin neurons

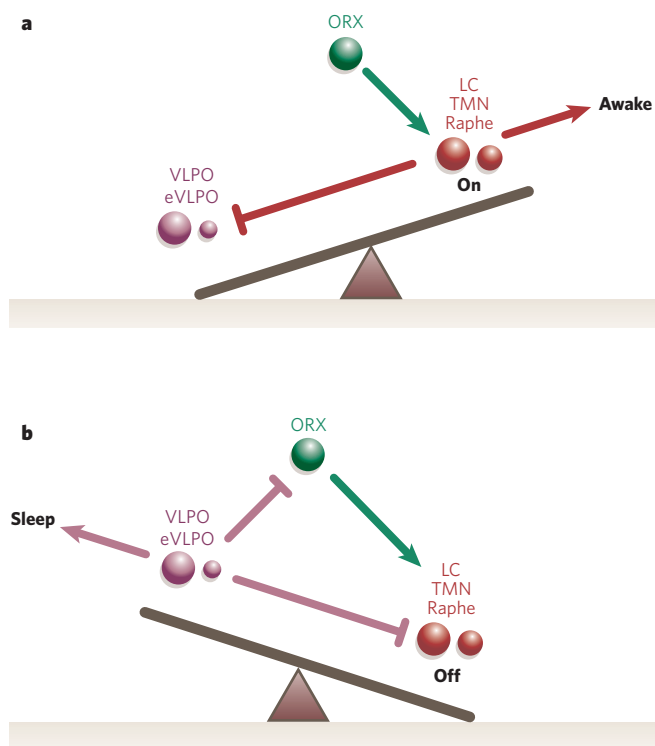


Figure 4 | A schematic diagram of the flip-flop switch model. During wakefulness (a), the monoaminergic nuclei (red) inhibit the ventrolateral preoptic nucleus (VLPO; purple), thereby relieving the inhibition of the monoaminergic cells, and that of the orexin (ORX) neurons (green), and the cholinergic pedunculopontine (PPT) and laterodorsal tegmental nuclei (LDT; yellow). Because the VLPO neurons do not have orexin receptors, the orexin neurons serve primarily to reinforce the monoaminergic tone, rather than directly inhibiting the VLPO on their own. During sleep (b), the firing of the VLPO neurons inhibits the monoaminergic cell groups, thereby relieving their own inhibition. This also allows it to inhibit the orexin neurons, further preventing monoaminergic activation that might interrupt sleep. The direct mutual inhibition between the VLPO and the monoaminergic cell groups forms a classic flip-flop switch, which produces sharp transitions in state, but is relatively unstable. The addition of the orexin neurons stabilizes the switch. 5-HT, serotonin; ACh, cholinergic; eVLPO, extended ventrolateral preoptic nucleus; GABA, γ -aminobutyric acid; gal, galanin; LC, locus coeruleus; NA, noradrenaline; PeF, perifornical; REM, rapid eye movement; TMN, tuberomammillary nucleus.

in the LHA and low orexin levels in the cerebrospinal fluid^{48–50}. So, within a remarkably short period of time, the molecular basis of this mysterious illness was identified.

Interestingly, few people with narcolepsy have mutations in either the orexin ligand or receptor genes^{49,51}. In most narcoleptics, the disease begins in the second or third decade of life, and the loss of orexin neurons is remarkably specific (producing no injury to the adjacent neurons that produce MCH)^{48,49}. The cause is believed to be autoimmune, although convincing evidence for this hypothesis is still lacking, and it might be a neurodegenerative condition⁵¹. Other patients with lesions of the posterior LHA due to a range of causes, such as those first reported by von Economo and Wilson, also acquired narcolepsy.

The orexin neurons are mainly active during wakefulness and especially during motor activity when animals actively explore their environment^{16–18}. They have ascending projections to the cerebral cortex, as well as descending projections to all the monoaminergic and cholinergic cell groups of the arousal systems^{47,52}. There are mutual projections between the VLPO neurons and the orexin neurons^{32,53,54}, but the former do not express orexin receptors⁵⁵. Hence, the orexin neurons reinforce the arousal systems, but probably do not directly inhibit the VLPO (Fig. 4). This asymmetric relationship could help stabilize the flip-flop switch, like a 'finger' on the switch that might prevent unwanted transitions into sleep. The increase in homeostatic sleep drive due to consolidated wakefulness might, in turn, help produce consolidated sleep. Narcoleptic people and animals lack this influence and behave as if their sleep flip-flop switch has been destabilized^{51,56}. They do not sleep more than normal individuals, but easily doze off during the day and wake more often from sleep at night⁵³, as the flip-flop model would predict.

Homeostatic regulation of sleep

Although the purpose of sleep remains unknown, it clearly has a restorative effect on the brain. Like other homeostatic systems, such as that governing body temperature, which seek to recover to a set point when perturbed, sleep deprivation is followed by extra recovery sleep that is proportional to the sleep loss. An influential model of sleep regulation, proposed by Borbely and colleagues, includes both homeostatic and circadian drives for sleep^{57,58}. The homeostatic influence is believed to be due to some structure or substance that accumulates 'need to sleep' during prolonged wakefulness, and discharges this

homeostatic need during sleep. NREM and REM sleep probably have separate homeostatic mechanisms, and, after a period of sleep deprivation, NREM sleep is typically repleted first.

The mechanisms for this homeostatic determinant remain unclear. The VLPO neurons, for example, do not accumulate a need to sleep, as they remain at a low level of activity after a prolonged period of sleep deprivation, until the animal actually falls asleep^{23,27}. At that point, VLPO neurons fire about twice as fast as they do during normal sleep²⁷, implying that they are under the influence of, but distinct from, the homeostatic factors that reflect sleep need.

Adenosine has been proposed as a homeostatic accumulator of the need to sleep^{6,59,60}. During prolonged wakefulness, the energy-producing systems in the brain run down, with exhaustion of brain glycogen reserves and depletion of ATP levels^{61,62}. During prolonged wakefulness, as ATP is degraded to ADP, AMP and eventually adenosine, extracellular adenosine levels rise in some parts of the brain, including the BF⁶³. Injection of adenosine or an adenosine A1 receptor agonist into the BF of cats⁶, or an adenosine A2a receptor agonist near the VLPO in rats, causes sleep; the latter also results in expression of Fos (a marker of neuronal activity) in VLPO neurons⁶⁴. In addition, adenosine might disinhibit the VLPO via presynaptic A1 receptors by reducing inhibitory GABAergic inputs³⁵. Hence, at least one mechanism for homeostatic sleep drive might be an accumulation of a sleep-promoting substance that enhances the activity of sleep-promoting cells and reduces the activity of wake-promoting neurons. In this way, adenosine, and perhaps other somnogens, can allow the activation of the VLPO that is necessary to trigger a sleep episode.

Circadian regulation of sleep

Borbely also proposed a circadian influence on sleep and wakefulness (process C) that is distinct from the homeostatic drive (process S) for sleep^{57,58}. Careful studies of humans in a forced desynchrony protocol, where they experience a 28-hour 'day' without any cues as to external time, have confirmed a strong 24-hour circadian rhythm in sleep drive⁶⁵. Recent studies in experimental animals have clarified how this drive is maintained.

The suprachiasmatic nucleus (SCN) serves as the brain's 'master clock'. Neurons in the SCN fire in a 24-hour cycle that is driven by a transcriptional–translational loop, which persists even when the neurons are dissociated in cell culture^{66,67}. Loss of the SCN abolishes the circadian rhythms of a range of behaviours and physiological

Box 1 | Drugs that affect sleep and wakefulness

Many of the drugs that are used to regulate sleep and wakefulness are now thought to act directly on the wake and sleep systems that have been outlined here.

Wake-promoting drugs, such as amphetamines, methylphenidate (Ritalin) and modafinil (Provigil), all act on the dopamine (DA)-reuptake transporter. Their potency is roughly proportional to their ability to block DA reuptake, and mice in which the DA-transporter gene has been deleted fail to respond to this entire class of medications⁹⁴. Modafinil might also interfere with noradrenaline reuptake, and this dual mode of action might explain why it can work without inducing other dopaminergic effects, such as addiction⁹⁵.

Some sleep-promoting drugs interfere with the actions of the ascending arousal system. For example, antihistamines that cross the blood–brain barrier, such as diphenhydramine (Benadryl), cause drowsiness by blocking the arousing influence of the central histaminergic system. This side effect accounts for the popularity of antihistamines that do not cross the blood–brain barrier and, hence, do not cause drowsiness. Centrally acting DA antagonists also cause drowsiness⁹⁶. These drugs are used primarily to treat psychotic disorders, and this side effect is often a limiting factor in drug acceptability. Paradoxically, D2-agonist drugs can also cause drowsiness at low dosages⁹⁶. The D2-presynaptic autoreceptor inhibits the firing of dopaminergic neurons, and therefore results in reduced dopaminergic activity. Similarly, dexmedetomidine, an α -2 adrenergic agonist, reduces firing of noradrenergic neurons, and is thought to induce loss of wakefulness by disinhibiting the ventrolateral preoptic nucleus

(VLPO), which is ordinarily under noradrenergic inhibition during wakefulness⁹⁷ (Fig. 4).

The largest class of sleep-promoting drugs consists of those that enhance the activity of γ -aminobutyric acid-A (GABA-A) receptors, including barbiturates, benzodiazepines, chloral hydrate, ethanol and most gaseous anaesthetics⁹⁸. Few of these drugs are direct agonists (that is, few compete with GABA for binding); rather, most bind to other components of the GABA-A receptor and enhance the action of GABA at the receptor. At low dosages, these drugs might act on the targets of the VLPO, which also uses GABA as a neurotransmitter to silence the arousal system. At higher dosages, they suppress firing in much of the central nervous system. Benzodiazepines, such as diazepam (Valium), act on GABA-A receptors that contain a γ -subunit and subunits of the α -1, α -2, α -3 or α -5 class⁹⁸. Newer 'non-benzodiazepines', such as zolpidem (Ambien) or eszopiclone (Lunesta), also bind at α - γ sites but do not have the classic benzodiazepine structure and are more selective for the α 1-containing receptors, which are important in sedation but not reduction of anxiety⁹⁸. Gaboxadol binds to GABA-A receptors that contain the α -4 and δ (but not γ) subunits. These receptors, which are at highest concentration in the thalamus, limbic system and cerebral cortex, are not accessible to benzodiazepines, and hence represent a distinct drug target⁹⁹. This might reduce the arousal drive from the medial temporal and prefrontal cortex that is seen during insomnia, as gaboxadol can activate Fos expression by neurons in the VLPO, and increases the amount of slow-wave sleep¹⁰⁰.

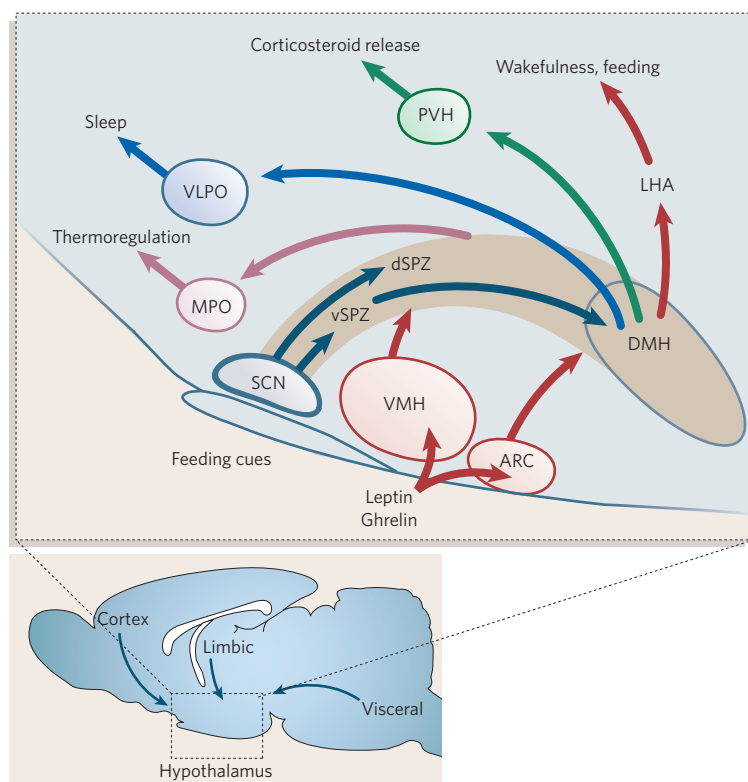


Figure 5 | A schematic diagram to illustrate the three-stage integrator for circadian rhythms. The suprachiasmatic nucleus (SCN) serves as a biological clock, but has few outputs to sleep-regulatory systems. Most of its output goes into the region in light brown, which includes the ventral (vSPZ) and dorsal (dSPZ) subparaventricular zone, and the dorsomedial nucleus of the hypothalamus (DMH). Neurons in the vSPZ relay information necessary for organizing daily cycles of wake-sleep, whereas dSPZ neurons are crucial for rhythms of body temperature. Outputs from the SPZ are integrated in the DMH with other inputs, and DMH neurons drive circadian cycles of sleep, activity, feeding and corticosteroid secretion. Cycles of body temperature are maintained by dSPZ projections back to the medial preoptic area (MPO), whereas the DMN is the origin of projections to the VLPO for sleep cycles, to the corticotropin-releasing hormone (CRH) neurons of the paraventricular nucleus (PVH) for corticosteroid cycles, and to the lateral hypothalamic (LHA) orexin and melanin-concentrating hormone neurons for wakefulness and feeding cycles. The integrative steps in the SPZ and DMH allow circadian rhythms to adapt to environmental stimuli, such as food availability (for example, leptin and ghrelin action via the ventromedial (VMH) and arcuate (ARC) nuclei), as well as visceral sensory inputs, cognitive influences from the prefrontal cortex and emotional inputs from the limbic system (inset).

processes, including sleep, if the animals are not given other external timing cues⁶⁸. Under normal circumstances, the SCN is reset on a daily basis by light inputs from the retina during the day, and by melatonin secretion from the pineal gland during the dark cycle^{69,70}. The light signal is received from a specialized set of retinal ganglion cells that contain the photopigment melanopsin⁷¹. These timing signals keep the clock in synchrony with the external day–night cycle.

The link between the SCN and the sleep system has been the subject of much recent investigation (Fig. 5). The SCN has relatively modest projections to the VLPO or the orexin neurons^{32,53,54,72}. However, the bulk of its output is directed toward the adjacent subparaventricular zone (SPZ) and the dorsomedial nucleus of the hypothalamus (DMH). The SPZ contains a ventral part, just above the SCN, and a dorsal part, just below the paraventricular nucleus⁷³. Cell-specific lesions of the ventral SPZ disrupt the circadian rhythms of sleep and wakefulness, as well as locomotor activity, but have minimal effects on body-temperature rhythms. Conversely, lesions of the dorsal SPZ severely impair circadian rhythms of body temperature, but not wake-sleep or locomotor activity⁷³. Therefore, direct projections from the SCN to sleep or thermoregulatory regions are not sufficiently strong to maintain the circadian rhythms of these functions, and the relay neurons in the SPZ are required.

The SPZ also has relatively limited projections to the VLPO, orexin neurons and other components of the wake-sleep-regulatory system^{32,53,54,74}. However, a major target is the DMH^{73,74}. This region receives inputs from many more neurons in the SPZ than the SCN, so the SPZ is in a position to amplify the output of the SCN. Cell-specific lesions of the DMH also profoundly diminish circadian rhythms of sleep and wakefulness, as well as locomotor activity, corticosteroid secretion and feeding⁷⁵. Interestingly, animals with DMH lesions sleep about one hour more each day and have much less locomotor activity, implying that the output of the DMH is mainly activating. This theme is also reflected in the corticosteroid levels, which, in animals with DMH lesions, remain at the lowest basal levels throughout the day. Body temperature retains a normal circadian variation, but is about 0.5 °C lower than in control animals⁷⁵.

The DMH is one of the largest sources of input to the VLPO and orexin neurons^{32,54,75,76}, and is crucial for conveying SCN influence to

the wake-sleep-regulatory system^{75,77}. The DMH projection to the VLPO comes largely from GABA-containing neurons (that is, those that promote wakefulness by inhibiting sleep), and the projection to the LHA originates from neurons containing glutamate and thyrotropin-releasing hormone (which should presumably be excitatory and promote wakefulness)⁷⁵. The DMH has relatively few direct outputs to the brainstem components of the ascending arousal system, but the orexin neurons have extensive projections to these targets^{47,52,76}. Examination of Fos patterns in the DMH show that it contains many more active neurons during wakefulness than during sleep⁷⁸.

Why might the brain need such a complex three-stage pathway for circadian control of sleep and other behaviours? The answer might lie in the observation that, in a wide range of both nocturnal and diurnal animals, the SCN is always active during the light cycle and the VLPO is always active during the sleep cycle^{23,24}. Therefore, in animals that are nocturnal versus diurnal, there must be an intervening set of circuitry that allows the circadian cycle to be set at opposite phases, despite an identical clock input and sleep-control system. In fact, the circadian rhythms in many animals are not entirely fixed. For example, bats are often considered to be the quintessential nocturnal animal. This is true for bats in Finland during the summer, when there are insects flying around at night for them to eat and predatory birds flying around during the day to discourage them. However, during the cooler weather of the spring and autumn, few insects fly at night and predatory birds have migrated to other climes. At these times of year, the activity cycles of bats shift so that they can take advantage of the food source that is available during the day, and they shift their activity pattern entirely to the daylight hours^{78,79}.

Similarly, it is possible to invert many of the circadian rhythms of laboratory rats by restricting their access to food to the daylight hours. For example, if animals are only permitted to eat during the latter half of the light period, they awaken, become active and have an increase in body temperature a few hours before the food is due to be presented⁸⁰. This behaviour persists even if the food is withheld completely for 2 days; hence, this circadian pattern is not dependent on external cues. The ability to alter the wake-sleep, activity, feeding, body temperature and corticosteroid rhythms of these animals correlates with a shift in activation of the DMH (but not the SCN, which

stays locked to the original light–dark cycle) to the new wake cycle, and lesions of the DMH prevent these behavioural shifts⁷⁸ (J. J. Gooley and C. B. Saper, unpublished observations). So, the DMH seems to integrate clock information from the SCN and SPZ with feeding, temperature, social and other cues, providing animals with the flexibility to adapt their behavioural and physiological cycles to the environment, thereby maximizing their chances of survival (Fig. 5).

Allostatic regulation of sleep

The rapid expansion of our knowledge about the substrates of the regulation of sleep–wakefulness and circadian rhythms highlights the lack of knowledge about how these systems are controlled in a complex and ever-changing environment. Although we are beginning to understand the homeostatic drive for sleep, it is clear that sleep patterns and the circadian cycle can be modified by external events, such as availability of food or external temperatures. Similarly, homeostatic and circadian drives for sleep can be overcome for brief periods when external events demand an emergency response.

McEwen and Stellar, in 1993, proposed the term ‘allostatic’ drive⁸¹ for situations where “rather than maintaining constancy, the physiologic systems within the body fluctuate to meet demands from external forces.” We know remarkably little about how these external forces overcome the homeostatic and circadian systems or reset them (Fig. 5). There are clearly cues from visceral sensory systems and feeding regulatory systems to the arousal systems^{54,82–85}. Visceral inputs relayed through the nucleus of the solitary tract, such as gastric stretch, have a synchronizing sleep-inducing influence⁸⁶, whereas absence of sufficient food causes an arousing influence⁸⁷. However, we know little about the mechanisms by which cognitive and emotional systems can act on the sleep or circadian control systems. Recent studies have emphasized the inputs to the SCN⁸², VLPO³² and orexin neurons^{53,54} from corticolimbic sites, such as the infralimbic cortex, ventral subiculum, lateral septum and bed nucleus of the stria terminalis, as possible routes by which this influence might be expressed.

Positron-emission tomography studies during sleep in humans with insomnia also show increased activity in corticolimbic sites, including the medial prefrontal cortex and medial temporal lobe, compared with sleeping subjects without insomnia⁸⁸. These inputs might maintain a hyperaroused state, which can be essential in an emergency situation, such as when a doctor must stay awake to care for a sick patient overnight. However, when the arousal systems override the homeostatic and circadian regulation of sleep during periods of behavioural stress or depression, the result might be unwanted and debilitating insomnia.

Conclusions

Understanding the interactions of these allostatic influences with the sleep and circadian systems is a major challenge for the next decade. However, learning to control our wake–sleep systems holds the promise of improved health and cognitive performance (Box 1). Individuals who lose even small amounts of sleep on a daily basis show progressive impairment of cognitive performance and elevation of C-reactive protein, which predicts cardiovascular risk^{89–91}. Because older individuals sleep about half an hour less per day, it is possible that at least some of their cognitive decline and increase in cardiovascular disease might be explained by sleep restriction⁹². Similarly, sleep loss might impair performance among adolescents who arise early for school, shift workers, overnight long-haul truckers and even medical personnel working in hospitals⁹³. The public-health implications of sleep loss indicate that there is a great deal at stake in working out the mechanisms that regulate our daily cycles of sleep and wakefulness. ■

1. Von Economo, C. Sleep as a problem of localization. *J. Nerv. Ment. Dis.* **71**, 249–259 (1930).
2. Moruzzi, G. & Magoun, H. W. Brain stem reticular formation and activation of the EEG. *Electroencephalogr. Clin. Neurophysiol.* **1**, 455–473 (1949).

3. Starzl, T. E., Taylor, C. W. & Magoun, H. W. Ascending conduction in reticular activating system, with special reference to the diencephalon. *J. Neurophysiol.* **14**, 461–477 (1951).
4. Saper, C. B., Chou, T. C. & Scammell, T. E. The sleep switch: hypothalamic control of sleep and wakefulness. *Trends Neurosci.* **24**, 726–731 (2001).
5. Hallanger, A. H., Levey, A. I., Lee, H. J., Rye, D. B. & Wainer, B. H. The origins of cholinergic and other subcortical afferents to the thalamus in the rat. *J. Comp. Neurol.* **262**, 104–124 (1987).
6. Strecker, R. E. *et al.* Adenosinergic modulation of basal forebrain and preoptic/anterior hypothalamic neuronal activity in the control of behavioral state. *Behav. Brain Res.* **115**, 183–204 (2000).
7. McCormick, D. A. Cholinergic and noradrenergic modulation of thalamocortical processing. *Trends Neurosci.* **12**, 215–220 (1989).
8. Krout, K. E., Belzer, R. E. & Loewy, A. D. Brainstem projections to midline and intralaminar thalamic nuclei of the rat. *J. Comp. Neurol.* **448**, 53–101 (2002).
9. Saper, C. B. Organization of cerebral cortical afferent systems in the rat. II. Hypothalamocortical projections. *J. Comp. Neurol.* **237**, 21 (1985).
10. Jones, B. E. Arousal systems. *Front. Biosci.* **8**, S438–S451 (2003).
11. Ranson, S. W. Somnolence caused by hypothalamic lesions in monkeys. *Arch. Neurol. Psychiatr.* **41**, 1–23 (1939).
12. Gerashchenko, D., Blanco-Centurion, C., Greco, M. A. & Shiromani, P. J. Effects of lateral hypothalamic lesion with the neurotoxin hypocretin-2-saporin on sleep in Long-Evans rats. *Neuroscience* **116**, 223–235 (2003).
13. Aston-Jones, G. & Bloom, F. E. Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep–waking cycle. *J. Neurosci.* **1**, 876–886 (1981).
14. Fornal, C., Auerbach, S. & Jacobs, B. L. Activity of serotonin-containing neurons in nucleus raphe magnus in freely moving cats. *Exp. Neurol.* **88**, 590–608 (1985).
15. Steininger, T. L., Alam, M. N., Gong, H., Szymusiak, R. & McGinty, D. Sleep–waking discharge of neurons in the posterior lateral hypothalamus of the albino rat. *Brain Res.* **840**, 138–147 (1999).
16. Estabrooke, I. V. *et al.* Fos expression in orexin neurons varies with behavioral state. *J. Neurosci.* **21**, 1656–1662 (2001).
17. Mileykovskiy, B. Y., Kiyashchenko, L. I. & Siegel, J. M. Behavioral correlates of activity in identified hypocretin/orexin neurons. *Neuron* **46**, 787–798 (2005).
18. Lee, M. G., Hassani, O. K. & Jones, B. E. Discharge of identified orexin/hypocretin neurons across the wake–sleep cycle. *J. Neurosci.* **25**, 6716–6720 (2005).
19. Verret, L. *et al.* A role of melanin-concentrating hormone producing neurons in the central regulation of paradoxical sleep. *BMC Neurosci.* **4**, 19 (2003).
20. Lee, M. G., Hassani, O. K., Alonso, A. & Jones, B. E. Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. *J. Neurosci.* **25**, 4365–4369 (2005).
21. Nauta, W. J. H. Hypothalamic regulation of sleep in rats. An experimental study. *J. Neurophysiol.* **9**, 285–314 (1946).
22. McGinty, D. J. & Sterman, M. B. Sleep suppression after basal forebrain lesions in the cat. *Science* **160**, 1253–1255 (1968).
23. Sherin, J. E., Shiromani, P. J., McCarley, R. W. & Saper, C. B. Activation of ventrolateral preoptic neurons during sleep. *Science* **271**, 216–219 (1996).
24. Gaus, S. E., Strecker, R. E., Tate, B. A., Parker, R. A. & Saper, C. B. Ventrolateral preoptic nucleus contains sleep-active, galaninergic neurons in multiple mammalian species. *Neuroscience* **115**, 285–294 (2002).
25. Sherin, J. E., Elmquist, J. K., Torrealba, F. & Saper, C. B. Innervation of histaminergic tuberomammillary neurons by GABAergic and galaninergic neurons in the ventrolateral preoptic nucleus of the rat. *J. Neurosci.* **18**, 4705–4721 (1998).
26. Szymusiak, R., Alam, N., Steininger, T. L. & McGinty, D. Sleep–waking discharge patterns of ventrolateral preoptic/anterior hypothalamic neurons in rats. *Brain Res.* **803**, 178–188 (1998).
27. Lu, J. *et al.* Selective activation of the extended ventrolateral preoptic nucleus during rapid eye movement sleep. *J. Neurosci.* **22**, 4568–4576 (2002).
28. Lu, J., Greco, M. A., Shiromani, P. & Saper, C. B. Effect of lesions of the ventrolateral preoptic nucleus on NREM and REM sleep. *J. Neurosci.* **20**, 3830–3842 (2000).
29. Verret, L. *et al.* Localization of the neurones potentially responsible for the inhibition of locus coeruleus noradrenergic neurones during paradoxical sleep in the rat. *J. Comp. Neurol.* (in press).
30. Ko, E. M., Estabrooke, I. V., McCarthy, M. & Scammell, T. E. Wake-related activity of tuberomammillary neurons in rats. *Brain Res.* **992**, 220–226 (2003).
31. John, J., Wu, M. F., Boehmer, L. N. & Siegel, J. M. Cataplexy-active neurons in the hypothalamus: implications for the role of histamine in sleep and waking behavior. *Neuron* **42**, 619–634 (2004).
32. Chou, T. C. *et al.* Afferents to the ventrolateral preoptic nucleus. *J. Neurosci.* **22**, 977–990 (2002).
33. Gallop, T. *et al.* Identification of sleep-promoting neurons *in vitro*. *Nature* **404**, 992–995 (2000).
34. Vincent, S. R., Hokfelt, T. & Wu, J.-Y. GABA neuron systems in hypothalamus and the pituitary gland. *Neuroendocrinology* **34**, 117–125 (1982).
35. Chamberlin, N. L. *et al.* Effects of adenosine on GABAergic synaptic inputs to identified ventrolateral preoptic neurons. *Neuroscience* **119**, 913–918 (2003).
36. Martin-Schild, S., Gerall, A. A., Kastin, A. J. & Zadina, J. E. Differential distribution of endomorphin 1- and endomorphin 2-like immunoreactivities in the CNS of the rodent. *J. Comp. Neurol.* **405**, 450–471 (1999).
37. Kohler, C. *et al.* Galanin immunoreactivity in hypothalamic neurons: further evidence for multiple chemical messengers in the tuberomammillary nucleus. *J. Comp. Neurol.* **250**, 58–64 (1986).
38. Chou, T. C. *Regulation of Wake–Sleep Timing: Circadian Rhythms and Bistability of Sleep–Wake States*, 82–99. Thesis, Harvard Univ. (2003).
39. Ancoli-Israel, S. & Kripke, D. F. Prevalent sleep problems in the aged. *Biofeedback Self Regul.* **16**, 349–359 (1991).
40. Hofman, M. A. & Swaab, D. F. The sexually dimorphic nucleus of the preoptic area in the human brain: a comparative morphometric study. *J. Anat.* **164**, 55–72 (1989).

41. Allen, L. S., Hines, M., Shryne, J. E. & Gorski, R. A. Two sexually dimorphic cell groups in the human brain. *J. Neurosci.* **9**, 497–506 (1989).
42. Byne, W. et al. The interstitial nuclei of the human anterior hypothalamus: an investigation of sexual variation in volume and cell size, number and density. *Brain Res.* **856**, 254–258 (2000).
43. Wilson, S. A. K. *Modern Problems in Neurology* (Edward Arnold, London, 1928).
44. Sakurai, T. et al. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
45. de Lecea, L. et al. The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proc. Natl Acad. Sci. USA* **95**, 322–327 (1998).
46. Lin, L. et al. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98**, 365–376 (1999).
47. Chemelli, R. M. et al. Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. *Cell* **98**, 437–451 (1999).
48. Thannickal, T. C. et al. Reduced number of hypocretin neurons in human narcolepsy. *Neuron* **27**, 469–474 (2000).
49. Peyron, C. et al. A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Med.* **6**, 991–997 (2000).
50. Ripley, B. et al. CSF hypocretin/orexin levels in narcolepsy and other neurological conditions. *Neurology* **57**, 2253–2258 (2001).
51. Scammell, T. E. The neurobiology, diagnosis, and treatment of narcolepsy. *Ann. Neurol.* **53**, 154–166 (2003).
52. Peyron, C. et al. Neurons containing hypocretin (orexin) project to multiple neuronal systems. *J. Neurosci.* **18**, 9996–10015 (1998).
53. Sakurai, T. et al. Input of orexin/hypocretin neurons revealed by a genetically encoded tracer in mice. *Neuron* **46**, 297–308 (2005).
54. Yoshida, K., McCormack, S., Espana, R. A., Crocker, A. D. & Scammell, T. E. Afferents to the orexin neurons. *J. Comp. Neurol.* (in the press).
55. Marcus, J. N. et al. Differential expression of orexin receptors 1 and 2 in the rat brain. *J. Comp. Neurol.* **435**, 6–25 (2001).
56. Mochizuki, T. et al. Behavioral state instability in orexin knock-out mice. *J. Neurosci.* **24**, 6291–6300 (2004).
57. Borbely, A. A. & Tobler, I. *Brain Mechanisms of Sleep* (eds McGinty, D. J. et al.) 35–44 (Raven, New York, 1985).
58. Achermann, P. & Borbely, A. A. Mathematical models of sleep regulation. *Front. Biosci.* **8**, S683–S693 (2003).
59. Radulovacki, M., Virus, R. M., Djuricic-Nedelson, M. & Green, R. D. Adenosine analogs and sleep in rats. *J. Pharmacol. Exp. Ther.* **228**, 268–274 (1984).
60. Benington, J. H. & Heller, H. C. Restoration of brain energy metabolism as the function of sleep. *Prog. Neurobiol.* **45**, 347–360 (1995).
61. Kong, J. et al. Brain glycogen decreases with increased periods of wakefulness: implications for homeostatic drive to sleep. *J. Neurosci.* **22**, 5581–5587 (2002).
62. Shepel, P. N., Ramonet, D., Stevens, P. & Geiger, J. D. Purine level regulation during energy depletion associated with graded excitatory stimulation in brain. *Neurol. Res.* **27**, 139–148 (2005).
63. Porkka-Heiskanen, T., Strecker, R. E. & McCarley, R. W. Brain site-specificity of extracellular adenosine concentration changes during sleep deprivation and spontaneous sleep: an in vivo microdialysis study. *Neuroscience* **99**, 507–517 (2000).
64. Scammell, T. E. et al. An adenosine A_{2a} agonist increases sleep and induces Fos in ventrolateral preoptic neurons. *Neuroscience* **107**, 653–663 (2001).
65. Dijk, D. J. & Czeisler, C. A. Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. *J. Neurosci.* **15**, 3526–3538 (1995).
66. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. *Nature* **418**, 935–941 (2002).
67. Jin, X. et al. A molecular mechanism regulating rhythmic output from the suprachiasmatic circadian clock. *Cell* **96**, 1–20 (1999).
68. Moore, R. Y. & Eichler, V. B. Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Res.* **42**, 201–206 (1972).
69. Johnson, R. F., Moore, R. Y. & Morin, L. P. Loss of entrainment and anatomical plasticity after lesions of the hamster retinohypothalamic tract. *Brain Res.* **460**, 297–313 (1988).
70. Cassone, V. M., Chesworth, M. J. & Armstrong, S. M. Entrainment of rat circadian rhythms by daily injection of melatonin depends upon the hypothalamic suprachiasmatic nuclei. *Physiol. Behav.* **36**, 1111–1121 (1986).
71. Gooley, J. J., Lu, J., Fischer, D. & Saper, C. B. A broad role for melanopsin in nonvisual photoreception. *J. Neurosci.* **23**, 7093–7106 (2003).
72. Watts, A. G., Swanson, L. W. & Sanchez-Watts, G. Efferent projections of the suprachiasmatic nucleus: I. Studies using anterograde transport of Phaseolus vulgaris leucoagglutinin in the rat. *J. Comp. Neurol.* **258**, 204–229 (1987).
73. Lu, J. et al. Contrasting effects of ibotenate lesions of the paraventricular nucleus and subparaventricular zone on sleep-wake cycle and temperature regulation. *J. Neurosci.* **21**, 4864–4874 (2001).
74. Deurveilher, S. & Semba, K. Indirect projections from the suprachiasmatic nucleus to major arousal-promoting cell groups in rat: implications for the circadian control of behavioural state. *Neuroscience* **130**, 165–183 (2005).
75. Chou, T. C. et al. Critical role of dorsomedial hypothalamic nucleus in a wide range of behavioral circadian rhythms. *J. Neurosci.* **23**, 10691–10702 (2003).
76. Thompson, R., Swanson, L. W. & Canteras, N. Organization of projections from the dorsomedial nucleus of the hypothalamus: a PHA-L study in the rat. *J. Comp. Neurol.* **376**, 143–173 (1997).
77. Aston-Jones, G., Chen, S., Zhu, Y. & Oshinsky, M. L. A neural circuit for circadian regulation of arousal. *Nature Neurosci.* **4**, 732–738 (2001).
78. Saper, C. B., Lu, J., Chou, T. C. & Gooley, J. The hypothalamic integrator for circadian rhythms. *Trends Neurosci.* **28**, 152–157 (2005).
79. Nyholm, H. Zur ökologie von *Myotis mystacinus* (Leisl.) und *M. daubentoni* (Leisl.) (Chiroptera). *Ann. Zool. Fenn.* **2**, 77–123 (1955).
80. Stephan, F. K. The “other” circadian system: food as a Zeitgeber. *J. Biol. Rhythms* **17**, 284–292 (2002).
81. McEwen, B. S. & Stellar, E. Stress and the individual. Mechanisms leading to disease. *Arch. Intern. Med.* **153**, 2093–2101 (1993).
82. Krout, K. E., Kawano, J., Mettenleiter, T. C. & Loewy, A. D. CNS inputs to the suprachiasmatic nucleus of the rat. *Neuroscience* **110**, 73–92 (2002).
83. Elmquist, J. K., Elias, C. F. & Saper, C. B. From lesions to leptin: hypothalamic control of food intake and body weight. *Neuron* **22**, 221–232 (1999).
84. Elmquist, J. K., Ahima, R. S., Elias, C. F., Flier, J. S. & Saper, C. B. Leptin activates distinct projections from the dorsomedial and ventromedial hypothalamic nuclei. *Proc. Natl Acad. Sci. USA* **95**, 741–746 (1998).
85. Saper, C. B. In *The Rat Nervous System* (ed. Paxinos, G.) 761–796 (Elsevier Academic, San Diego, 2004).
86. Schachter, S. C. & Saper, C. B. Vagus nerve stimulation. *Epilepsia* **39**, 677–686 (1998).
87. Yamanaka, A. et al. Hypothalamic orexin neurons regulate arousal according to energy balance in mice. *Neuron* **38**, 701–713 (2003).
88. Nofzinger, E. A. et al. Functional neuroimaging evidence for hyperarousal in insomnia. *Am. J. Psychiatry* **161**, 2126–2128 (2004).
89. Meier-Ewert, H. K. et al. Effect of sleep loss on C-reactive protein, an inflammatory marker of cardiovascular risk. *J. Am. Coll. Cardiol.* **43**, 678–683 (2004).
90. Van Dongen, H. P., Maislin, G., Mullington, J. M. & Dinges, D. F. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep* **26**, 117–126 (2003).
91. Doran, S. M., Van Dongen, H. P. & Dinges, D. F. Sustained attention performance during sleep deprivation: evidence of state instability. *Arch. Ital. Biol.* **139**, 253–267 (2001).
92. Prinz, P. N. Age impairments in sleep, metabolic, and immune functions. *Exp. Gerontol.* **39**, 1739–1743 (2004).
93. Landrigan, C. P. et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N. Engl. J. Med.* **351**, 1838–1848 (2004).
94. Wisor, J. et al. Dopaminergic role in stimulant-induced wakefulness. *J. Neurosci.* **21**, 1787–1895 (2001).
95. Saper, C. B. & Scammell, T. E. Modafinil: a drug in search of a mechanism. *Sleep* **27**, 11–12 (2004).
96. Roth, T. et al. Assessment of sleepiness and unintended sleep in Parkinson's disease patients taking dopamine agonists. *Sleep Med.* **4**, 275–280 (2003).
97. Nelson, L. E. et al. The 4-2 adrenoreceptor agonist dexmedetomidine converges on an endogenous sleep-promoting pathway to exert its sedative effects. *Anesthesiology* **98**, 428–436 (2003).
98. Rudolf, U. & Mohler, H. Analysis of GABAA receptor function and dissection of the pharmacology of benzodiazepines and general anesthetics through mouse genetics. *Annu. Rev. Pharmacol. Toxicol.* **44**, 475–498 (2004).
99. Wisden, W., Laurie, D. J., Monyer, H. & Seeburg, P. The distribution of 13 GABAA receptor subunit mRNAs in the rat brain. I. Telencephalon, diencephalon, mesencephalon. *J. Neurosci.* **12**, 1040–1062 (1992).
100. Lancel, M., Wetter, T. C., Steiger, A. & Mathias, S. Effect of the GABAA agonist gaboxadol on nocturnal sleep and hormone secretion in healthy elderly subjects. *Am. J. Physiol. Endocrinol. Metab.* **281**, E130–E137 (2001).

Acknowledgements This work was supported by United States Public Health Service grants.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence should be addressed to C.F.B. (csaper@bidmc.harvard.edu).

Clues to the functions of mammalian sleep

Jerome M. Siegel¹

The functions of mammalian sleep remain unclear. Most theories suggest a role for non-rapid eye movement (NREM) sleep in energy conservation and in nervous system recuperation. Theories of REM sleep have suggested a role for this state in periodic brain activation during sleep, in localized recuperative processes and in emotional regulation. Across mammals, the amount and nature of sleep are correlated with age, body size and ecological variables, such as whether the animals live in a terrestrial or an aquatic environment, their diet and the safety of their sleeping site. Sleep may be an efficient time for the completion of a number of functions, but variations in sleep expression indicate that these functions may differ across species.

Saying that it is desirable to be well rested and that the body seeks lost sleep with a vigour comparable to or greater than that displayed for food or sex does not answer the question of the functional role of sleep. Why do we spend one-third of our lives asleep? Why has our body evolved to press us relentlessly to make up for lost sleep? Can we separate the drive for sleep, manifested in sleepiness, from the function of sleep, as we can separate hunger from the benefits of food consumption? Why do so many species habitually sleep much more than humans, and others much less, and how do species that sleep for only short periods accomplish the functions of sleep in less time? Why does the daily sleep amount decrease from birth to maturity in all species of terrestrial mammals? And why do we have two kinds of sleep, rapid eye movement (REM) and non-REM (NREM) sleep?

Sleep can be defined as a state of immobility with greatly reduced responsiveness, which can be distinguished from coma or anaesthesia by its rapid reversibility. An additional defining characteristic of sleep is that when it is prevented, the body tries to recover the lost amount. The existence of sleep 'rebound' after deprivation¹ demonstrates that sleep is not simply a period of reduced activity or alertness regulated by circadian or ultradian rhythms, a phenomenon that can be seen even in non-sleeping organisms²⁻⁴.

The amplitude of the changes in brain metabolism and neuronal activity that occurs during sleep exceeds those which occur during most waking periods⁵⁻⁷. The argument that sleep serves a vital function is compelling. Sleep deprivation in rodents and flies can cause death more quickly than food deprivation⁸. Nevertheless, we must not assume that the effects of sleep loss are independent of the deprivation technique used or that sleep loss has equally dire effects in all animals^{9,10}.

In this review we will consider the vast knowledge that has been gained about the physiological nature of sleep and sleep-control mechanisms, evidence from sleep-deprivation studies and the distribution of sleep across species in the context of theories of sleep function. These data support theories that suggest that sleep saves energy, keeps species from being active at inopportune times and reverses waking-induced changes in brain function. The evidence suggests distinct roles for REM and NREM sleep. It is also clear that sleep expression is adapted to ecological factors and may differ qualitatively across species.

Sleep-controlling brain regions in mammals

Neurophysiological studies have provided considerable information about the mechanisms controlling sleep states. These data can guide

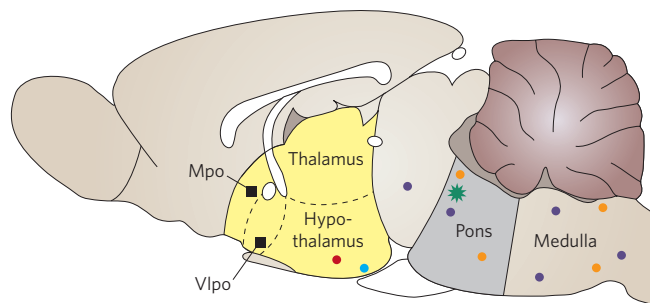


Figure 1 | Distribution of some key sleep-regulating neuronal populations plotted on a sagittal section of a rat brain⁹⁷. Circles indicate 'REM sleep off' neurons; purple represents serotonergic neurons (located on the midline), orange represents adrenergic or noradrenergic neurons, blue represents histaminergic neurons, red represents hypocretinergic (orexinergic) neurons. Squares indicate 'sleep on' neurons. The green star indicates 'REM sleep on' neurons. The area shaded in grey is both necessary and sufficient for REM sleep generation. The area shaded in yellow is both necessary and sufficient for NREM sleep generation. In the intact animal both REM sleep and NREM sleep involve interactions between brainstem and forebrain structures. Vlo, ventrolateral preoptic area; Mpo, median preoptic.

theories of sleep functions. Detailed reviews of the physiological control of sleep are available elsewhere¹¹, but for the purposes of the current review, several aspects will be highlighted.

NREM sleep phenomena can be generated by the isolated forebrain¹²⁻¹⁴. Groups of sleep-active neurons have been discovered in the preoptic and basal forebrain regions (Fig. 1). These cells are maximally active during NREM sleep, and when stimulated will induce this state. Conversely, damage to these regions greatly reduces sleep. These neurons act through direct and indirect inhibitory projections to aminergic, cholinergic and hypocretinergic (also called orexinergic) neurons in the forebrain and brainstem. These and other neuronal groups maintain waking. The preoptic and anterior hypothalamic regions, within which most of these sleep-active neurons are embedded, have central roles in controlling the body and brain's temperature¹⁵. Many sleep-active neurons are thermosensitive; when studied in tissue slices and in the intact brain they increase their activity at higher temperatures¹². Heating of the preoptic regions increases NREM sleep.

REM sleep phenomena can be generated by the isolated brainstem,

¹Neurobiology Research 151A3, VA GLAHS Sepulveda, Department of Psychiatry and Brain Research Institute, UCLA School of Medicine, North Hills, California 91343, USA.

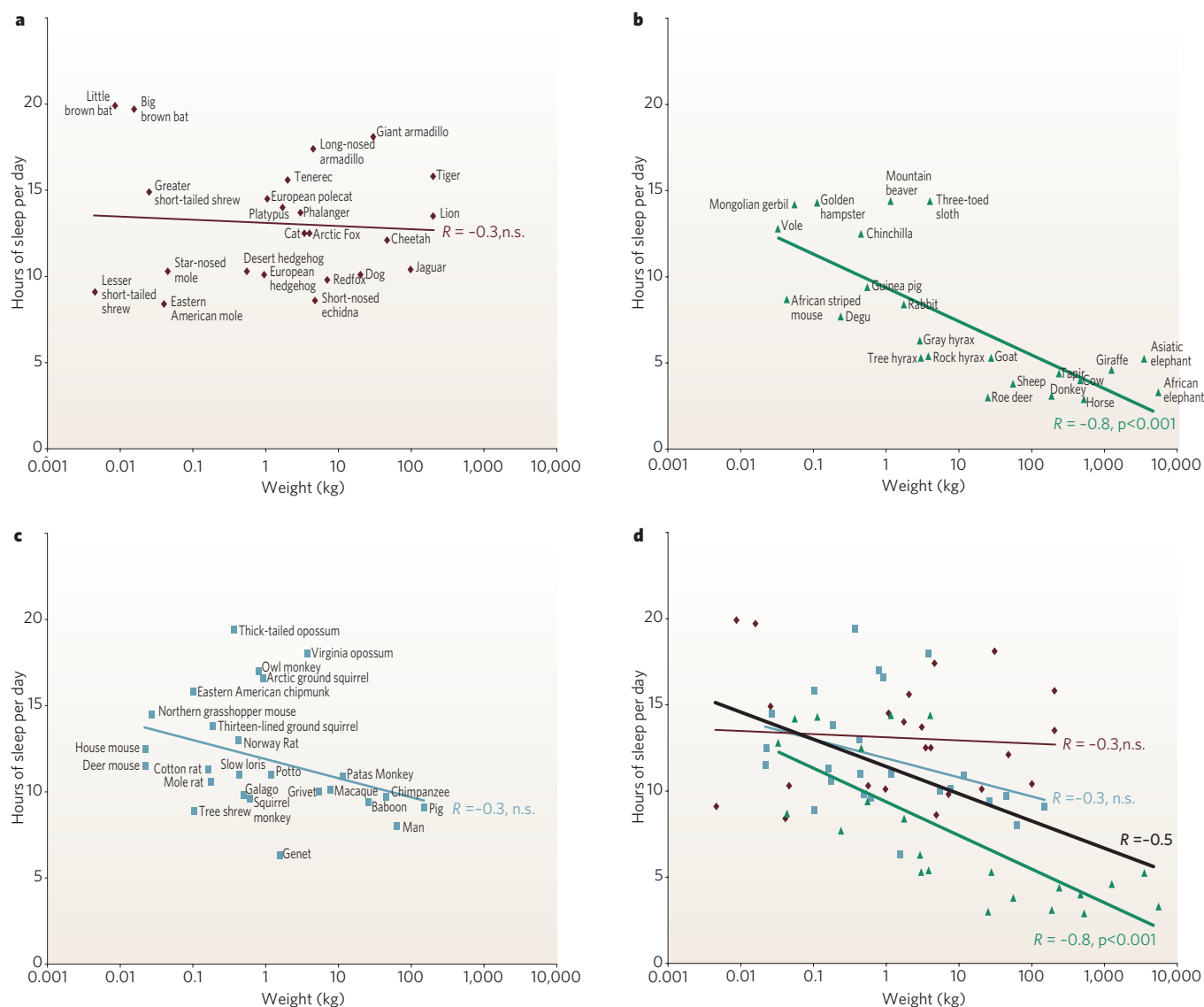


Figure 2 | Sleep time in mammals. **a**, Carnivores are shown in dark red; **b**, herbivores are in green and **c**, omnivores in grey. Sleep times in carnivores, omnivores and herbivores differ significantly ($P < 0.0002$, F test, d.f. 2,68), with carnivore sleep amounts significantly greater than those of herbivores ($P < 0.001$, t -test, d.f. 24, 22). Sleep amount is an inverse function of body mass over all terrestrial mammals (black line). This function accounts for approximately 25% of the interspecies variance (**d**) in reported sleep amounts (regression of log weight against sleep amount, $R = -0.5$, $P < 0.0001$, $n = 71$). Herbivores are responsible for this relation because body mass and sleep time were significantly and inversely correlated in herbivores ($R = -0.77$, $P < 0.001$, d.f. 24), but were not in carnivores ($R = -0.28$, d.f. 24) or omnivores ($R = -0.25$, d.f. 25).

It is not easy to quantify sleep parameters throughout the animal kingdom and as a result all desired parameters are rarely measured. For example, arousal thresholds are rarely systematically measured as part of a phylogenetic sleep study. Evidence for homeostatic regulation of sleep (sleep rebound) is seldom sought. Most species have not been implanted with electrodes for monitoring muscle tone and other variables. Instead, estimates are often based on visual observations, with the observer forced to intuit the differences between quiet waking and sleep. Other factors such as temperature, light cycle, food and noise conditions, which all affect sleep, have often not been controlled for. The age and health of the animals observed can vary, particularly depending on whether observations are made on animals in the wild or in the zoo. Often, observations of only one or two individuals are the source of the reported sleep amount of a given species. In animals observed in the wild, the weight of the subject is often not known, and in many cases the typical adult body weight, brain weight and other anatomical and physiological parameters cannot be or have not been precisely determined. Despite these sources of noise, significant relationships between weight, sleep time and diet are apparent. Source data for this figure were mainly from ref. 25.

specifically by the pons and adjacent midbrain¹⁴. This region contains a subgroup of neurons that are maximally active during REM sleep. These 'REM sleep on' cells cause a complete loss of muscle tone in the postural muscles during REM sleep by triggering simultaneous inhibition of and withdrawal of excitation to motoneurons. They also have a crucial role in the regulation of REM sleep itself. When these neurons are activated, which can be accomplished by microinjecting acetylcholine agonists into specific regions of the pons, prolonged REM sleep periods are elicited¹⁶. Damage to these neurons greatly diminishes or prevents REM sleep for long periods¹⁴. Although the ther-

mosensitivity of the REM sleep on cell populations has not been characterized, it has been shown that cooling of the isolated brainstem produces a marked increase in REM sleep amount¹⁷, just as heating of the forebrain produces a marked increase in NREM sleep¹². This suggests that, in the intact animal, brainstem mechanisms triggering REM sleep may be facilitated by brainstem cooling or correlated metabolic changes.

The effects of localized temperature changes in brainstem and forebrain regions on sleep should not be confused with the effects of environmental temperature. Sleep is reduced at temperatures outside the

thermoneutral zone, and REM sleep amounts are maximal at the upper levels within this zone¹⁵.

Neuronal activity across the sleep cycle

With the important exception of the comparatively small populations of sleep-active and REM sleep-active neurons described above, and other small groups of 'REM sleep off' neurons to be considered below, most brainstem neurons are maximally active during waking with movement and during REM sleep and are minimally active during NREM sleep¹⁴. In neocortical neurons, the decrease in activity during NREM sleep compared with active waking is smaller than that seen in the brainstem, but the neuronal activity pattern differs greatly from the REM sleep and waking pattern. The asynchronous activity of adjacent neocortical neurons in both REM sleep and waking changes to a rhythmic discharge, synchronized across large regions of the neocortex. This causes a summation of excitatory and inhibitory postsynaptic potentials, resulting in the high-voltage 2–12-Hz neocortical waves that are seen by electroencephalogram (EEG)¹⁸. These activity patterns are linked to maximal metabolic activity during waking and REM sleep, and minimal metabolic activity during NREM sleep in both the neocortex and the brainstem^{6,7}.

If cortical and subcortical neuronal activity is similar during waking and REM sleep, why are these states so different? The differences between these two states, including the difference in muscle tone¹⁹ and the different levels of awareness of the environment, can presumably be attributed to the small number of neurons whose activity differs between these two states. Two groups of cells whose activity differs between waking and REM sleep are the REM on cells, discussed above, and the REM sleep off cells. The latter cells, which contain noradrenaline, epinephrine, serotonin, histamine or hypocretin, are continuously active during waking. They all have decreased activity during NREM sleep (this decrease is particularly marked for hypocretin cells, which may also be silent during quiet waking) and cease discharge during REM sleep^{14,20–23}. Recent work suggests that the cessation of histamine neuron activity is linked to the loss of consciousness in sleep, whereas the cessation of noradrenergic neuron activity is linked to muscle tone suppression during sleep²⁴.

Sleep studies in terrestrial mammals

To evaluate theories of REM and NREM sleep function, one must consider how sleep amounts differ across species. Daily sleep amounts vary substantially from mammal to mammal. Some animals, such as bats and opossums, sleep for 18–20 hours a day (Fig. 2). Others, such as the elephant and giraffe, sleep for as little as 3–4 hours a day. One might expect species in each mammalian order to have a similar sleep pattern because of their genetic, behavioural and anatomical similarities. This is not the case. Differences in order do not simply explain differences in sleep amounts²⁵. Primates as a group do not have sleep characteristics that distinguish them from Rodentia, Insectivora or other orders. Humans, in particular, do not seem to have amounts or aspects of REM sleep or NREM sleep that distinguish them from other species, although they do have less sleep (more waking) than most omnivores (Fig. 2).

A species' customary diet is correlated with sleep time. Daily sleep amounts are highest in carnivores, lower in omnivores and lowest in herbivores. Sleep time is inversely correlated with body mass in herbivores. This correlation is responsible for a significant overall correlation between body mass and sleep time in all mammals studied so far²⁵; however, sleep time and body mass are not significantly correlated in carnivores or omnivores when they are evaluated separately (Fig. 2).

Another aspect of sleep strongly linked to body mass and brain size is the duration of the sleep cycle, that is, the average time taken to cycle from NREM sleep onset, through REM sleep, to waking. Small animals have shorter sleep cycles, and cycle times range from about 8 minutes in the short-tailed shrew (*Blarina brevicauda*) to 1.8 hours in the Asiatic elephant (*Elephas maximus*)²⁵. The reason for this robust correla-

tion is not known. Possible explanations include the inverse correlation of metabolic rate with body mass and brain mass, the thermal inertia of the brain and body, the time required for diffusion of substances through the brain parenchyma or the time required to complete a particular anabolic or catabolic biochemical task.

Most studies of mammalian sleep have been performed on placental (eutherian) or marsupial mammals. The third subclass of mammals is the monotremes, found in Australia and New Guinea. These egg-laying mammals have more genetic and physiological similarities to reptiles and birds than do other mammals and are thought to possess characteristics of the common mammalian ancestor²⁶. Although an initial study of the echidna suggested that monotremes do not have REM sleep²⁷, further observations indicated that they have an unusual form of this state. Both the echidna and platypus show evidence of brainstem activation during sleep, with the platypus displaying intense rapid eye, limb and bill movements periodically during sleep. However, the low-voltage neocortical EEG typically seen in eutherian and marsupial mammals during REM sleep is not consistently present during sleep in either the echidna or platypus during these activities^{28,29}. Thus, these 'primitive' mammals seem to have a form of REM sleep that is mainly localized to the brainstem.

Sleep in marine mammals

All terrestrial mammals show relatively high-voltage neocortical EEG activity bilaterally during NREM sleep. By contrast, cetaceans (whales and dolphins) almost never have high-voltage slow waves in both hemispheres at the same time (Fig. 3; refs 30, 31). Manatees (*Trichechus inunguis*, a member of the order Sirenia) also have uni-hemispheric slow waves³². In all marine mammals studied to date, the eye contralateral to the brain hemisphere with slow waves is almost always closed while the other eye is almost always open. There have been no published reports documenting REM sleep in cetaceans, making them the only studied mammals in which this state has not been observed.

The bottlenose dolphin (*Tursiops truncatus*), when not floating or resting on the bottom, generally swims in a single direction (usually counterclockwise) even as the hemisphere with slow waves alternates. Some smaller cetacean species are rarely, if ever, immobile. They move and avoid obstacles 24 hours a day from birth until death, even during uni-hemispheric slow-wave activity. These animals may never exhibit the immobility that we use in terrestrial mammals to define the state of sleep³³.

Slow (high-voltage 1–4-Hz) waves and spindle (high-voltage 8–12 Hz) waves are not by themselves conclusive evidence for sleep as defined at the beginning of this review. They have been linked to sleep because they generally accompany the behavioural signs of sleep in terrestrial mammals. However, even in terrestrial mammals, certain drugs can induce bilateral high-voltage EEG activity in individuals that are clearly awake³⁴. Moreover, high-voltage neocortical activity is not a required indicator of sleep; a NREM sleep state with low-voltage neocortical activity has been documented in rodents³⁵. REM sleep is characterized by a low-voltage EEG, indistinguishable from that of waking in many animals.

Studies of arousal threshold have not been performed on cetaceans across putative sleep–wake cycles. Only one study has looked for sleep rebound after EEG slow waves were prevented by disturbing cetaceans. Although some evidence for rebound was seen, the response was highly variable, with some animals showing little or no recovery of lost slow waves. The amount of slow waves after deprivation bore no consistent relation to the amount of slow-wave activity that had been lost³⁶. Neuronal recording studies in terrestrial mammals suggest that the axial movements occurring in cetaceans during both normal swimming and swimming with uni-hemispheric slow waves are accompanied by activation of large regions of the brainstem reticular formation. This is quite different from the great reduction in brainstem reticular activity that characterizes NREM sleep in terrestrial mammals^{14,37}. Cetaceans deftly avoid obstacles during this constant motion, sug-

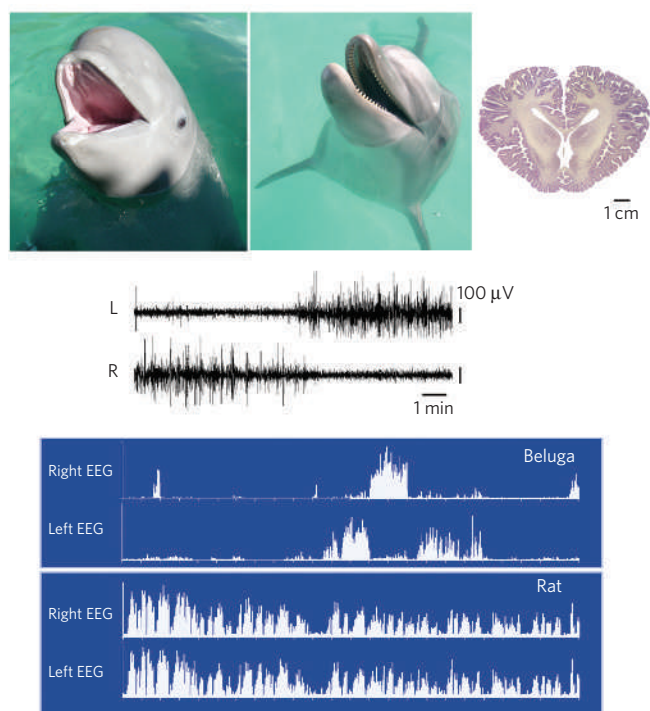


Figure 3 | Unihemispheric slow waves in cetaceans. Top, photos of immature beluga, adult dolphin and section of adult dolphin brain. Electroencephalogram (EEG) of adult cetaceans, represented here by the beluga, during sleep are shown. All species of cetacean so far recorded have unihemispheric slow waves^{30,31,98–100}. Top traces show left and right EEG activity. The spectral plots show 1–3-Hz power in the two hemispheres over a 12-hour period. The pattern in the cetaceans contrasts with the bilateral pattern of slow waves seen under normal conditions in all terrestrial mammals, represented here by the rat (bottom traces). The brain photograph is from the University of Wisconsin, Michigan State, and the National Museum of Health Comparative Mammalian Brain Collections.

gesting accurate bilateral processing of sensory information.

If sensory and motor systems do not show typical sleep inactivation, if behavioural and sleep rebound evidence for sleep debt is weak, do cetaceans sleep as conventionally defined? Certainly further work is necessary to determine which, if any, neurochemical and neurophysiological aspects of sleep, other than unilateral neocortical slow waves and eye closure, are preserved in cetaceans and might constitute the 'essence' of sleep. Alternatively, we may need to revise our assumption that sleep is fundamentally similar in cetaceans and terrestrial mammals and focus on how cetaceans can dispense with some of the most readily detectable aspects of sleep.

Postpartum sleep behaviour in cetaceans

Further evidence for the unique properties of 'sleep' in cetaceans is the near absence of sleep behaviour in neonates and a postpartum reduction in sleep behaviour in their mothers³⁸. All terrestrial mammals have minimal activity and maximal total sleep and REM sleep amounts at birth, with sleep gradually decreasing and activity gradually increasing to adult levels as the animals grow to maturity³⁹. This is not the pattern in cetaceans. We have found that killer whales (*Orcinus orca*) and dolphins have minimal amounts of sleep behaviour (that is, immobility or eye closure) at birth, with sleep behaviour slowly increasing to adult levels over a period of months. This minimal amount of sleep behaviour occurs during the period of most rapid growth of the body and brain for the newborn, during a period of bonding to the mother and learning how to nurse, find food, avoid predators and swim efficiently.

We have hypothesized that the continuous activity of cetaceans has adaptive value in allowing the neonate, which is much less insulated by body fat than the adult, to thermoregulate in cold ocean water. The

suppression of sleep behaviour also allows the neonate to swim with and be protected by its mother during development. Unlike the terrestrial environment, there are few warm, safe places to sleep in the ocean. As the animal gains mass and blubber and approaches adult size, adult-like 'sleep' or rest behaviour, including periods of immobility, emerges.

Equally striking is the near absence of typical sleep behaviour in the mother during the postpartum period. Both mother and calf go without substantial amounts of immobility and the eye closure linked to unihemispheric slow waves, for periods that greatly exceed those of sleep deprivation that are reported to kill rats⁸. Moreover, neither the mother nor the calf shows any rebound increase in the amount of sleep behaviour after this period.

The suppression of sleep behaviour seen in cetaceans after birth is analogous to a reduction in sleep that has been seen during the migration season in one bird species studied, the white crowned sparrow. This sleep reduction was accompanied by high levels of alertness and was not followed by sleep rebound. It has been argued that this phenomenon may be similar to mania seen in humans, which is also accompanied by a major reduction in sleep⁴⁰.

Fur seals

Fur seals and other otariids (seals with external ear flaps) exhibit another difference from the sleep of terrestrial mammals. When in the water, fur seals show a pattern of unihemispheric slow waves similar to that of cetaceans. However, fur seal sleep, unlike that of cetaceans, is accompanied by striking motor asymmetries. The flipper contralateral to the hemisphere with slow waves is immobile, the flipper contralateral to the hemisphere with low-voltage activity paddles to maintain the animal's position, and the whisker contralateral to the hemisphere with low-voltage activity is used to monitor the seal's position in the water. When slow-wave activity shifts to the opposite side, so do the motor asymmetries. When in the water, the fur seal has an extremely small amount of REM sleep, and may go without any REM sleep for a week or two. Surprisingly, when the fur seal moves onto land after spending weeks in the water, a change in its sleep structure occurs immediately. Unihemispheric slow waves largely disappear, and bilateral NREM and REM sleep are present, similar in amount and form to those of comparably sized land mammals. A remarkable feature of the transition from water to land is that fur seals immediately adopt their land sleep amounts without any evidence of REM sleep rebound, despite the absence or near absence of REM sleep while in the water⁴¹.

Theories of NREM sleep function

There is no shortage of theories to explain the functions of REM and NREM sleep. Recent reviews have critically evaluated many of these theories^{8,15,42,43}. There are several themes that can be seen in some of these theories and for which the data reviewed above provide support, and other theories that are inconsistent with these data. In general, most theories have assumed that sleep serves the same function in all animals, although most data supporting such theories have been derived from observations limited to just a few species of terrestrial mammal.

Neocortical maintenance

A recurrent theme in theories of sleep function is that sleep time is determined by neuronal activity in the neocortex. Although neocortical EEG changes are the most easily observed electrical correlate of sleep because they are recordable from scalp electrodes in humans and from electrodes placed on the surface of the cortex in other animals, sleep produces large changes in the rates and patterns of neuronal activity in nearly all brain regions, not just in the neocortex. Cortical EEG phenomena are controlled by and reflect activity in thalamic, hypothalamic and brainstem reticular regions. The cellular activity changes underlying the changes in neocortical EEG include calcium fluxes into and hyperpolarization of neocortical and thalamic neurons that are synchronized in large populations, producing high-voltage

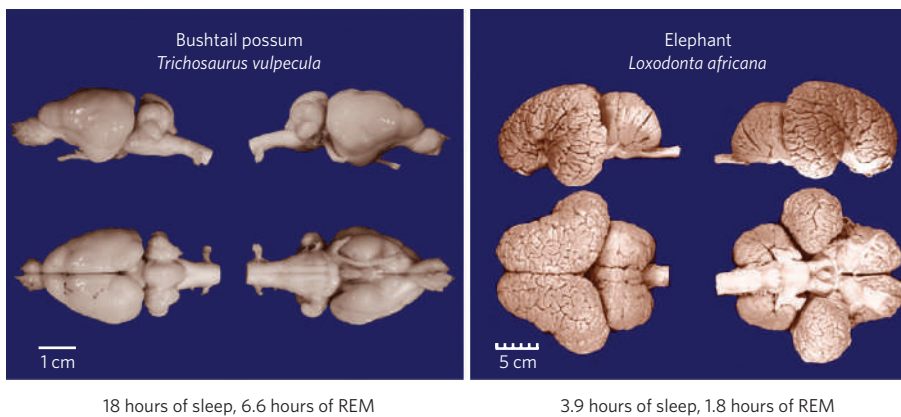


Figure 4 | Size of the neocortex does not correlate positively with daily sleep amount. Sleep amount is not proportional to the relative size of the cerebral cortex or to the degree of encephalization, as illustrated by these two examples. Brain photographs are from the University of Wisconsin, Michigan State, and the National Museum of Health Comparative Mammalian Brain Collections.

brain waves^{18,44}. However, neocortical size does not correlate positively with sleep amount. Both total brain weight and encephalization correlate poorly and negatively with total NREM and REM sleep amounts²⁵. The elephant, which has the largest neocortex of any terrestrial mammal, has one of the smallest amounts of sleep. Conversely, the rat and the platypus, which have smooth cortices with small total neocortical volumes, have extremely large amounts of sleep, with the platypus having more REM sleep than any other animal studied so far (Fig. 4; ref. 29).

Although neocortical size does not seem to be a major determinant of either NREM or REM sleep amounts, recent work has indicated that neocortical activity during sleep may be altered by prior waking activity. Some such changes dissipate with continued waking^{44–46}, suggesting that localized recuperative processes may take place during either waking or sleep in systems projecting to, or within, the neocortex.

Effects of sleep deprivation

Sleep restriction leads to a feeling of sleepiness and, depending on the nature of the sleep lost, to increases in the amplitude of the brain-wave signals that characterize NREM sleep and the amplitudes and frequencies of the eye movements and twitches that characterize REM sleep, when sleep is allowed^{47,48}. Sleep loss causes intrusions of sleep into waking that can displace behaviours that have obvious survival value. The fact that sleep debt can be accumulated suggests that sleep serves important functions that require some portion of the missed sleep amount to be made up. Although motivated humans can overcome sleepiness for short periods, they cannot perform at high levels for sustained periods^{49,50}.

The signs of long-term sleep deprivation, including skin lesions, hyperthermia followed by hypothermia, increased food intake and death, that have been noted in the rat, the subject whose response has been most thoroughly investigated, do not occur in the rat or cat even with total, long-term decortication⁵¹, consistent with the lack of a positive correlation between cortical size and sleep time noted above. Many of these signs of sleep loss can be seen after hypothalamic damage and concomitant abnormal functioning of the endocrine and immune systems⁵².

Energy conservation

Sleep may be adaptive because it conserves energy and suppresses behaviour across portions of the 24-hour day, just as hibernation does across certain seasons⁵³. Large herbivores may have evolved reduced sleep amounts because they are more vulnerable to predators than small herbivores⁵⁴. A second hypothesis is that these grazing animals may need to spend more time awake in order to eat because of the low caloric density of their food. A complementary hypothesis is that small herbivores and other mammals may need to maximize sleep amounts to conserve energy because their relatively high ratio of surface area to body mass makes it costly to maintain their body temperature. Retreating to a warm, protected nest may minimize this cost. A striking feature of sleep in animals with small daily sleep amounts,

such as many herbivores, is that sleep depth, as judged by EEG and sensory response threshold, seems to be less than that in animals requiring more sleep. In other words, animals with low sleep amounts do not seem to 'make up' for low sleep quantity by sleeping more 'deeply'. This contrasts with the raised arousal thresholds and increased amplitudes of low-frequency brain-wave activity shown in individual animals after sleep deprivation^{55,56}.

Energy conservation may be particularly important in newborns. Their high ratio of surface area to body mass makes the energy conservation achieved by sleep highly adaptive. Furthermore, animals that are immature at birth benefit from the sleep-induced reduction in exposure to danger. When body size increases and sensory-motor systems mature, young animals derive greater benefits from waking activities, consistent with the developmental decrease in sleep time.

Body mass, metabolism and sleep control

One of the best-established relationships in mammalian biology is the inverse link between body mass and mass-specific metabolic rate. Small animals have high metabolic rates; large animals have low metabolic rates. Brain metabolic rate is correlated with body metabolic rate⁵⁷. Elevated metabolism is linked to a number of biochemical changes, several of which have been linked to sleep control.

Sleep time may be related to defence against oxidative stress, particularly in, but not limited to, herbivores. A high metabolic rate results in the generation of high levels of reactive oxygen species (ROS) by mitochondria. This ROS generation has been linked to ageing, producing a wrinkled, arthritic, demented mouse by two years of age, whereas larger animals such as humans do not experience such effects of ageing until they are 70–80 years old. Previous studies challenging the phylogenetic relationship between sleep time and metabolic rate have used the partial correlation statistical approach. This statistical procedure is confounded when used with highly correlated variables such as metabolic rate and body mass.

We and others have shown that sleep deprivation in the rat is accompanied by increased oxidative stress and evidence of membrane disruption in the hippocampus, subcortical brain regions and peripheral tissues^{58–60}. There were no such changes in the neocortex^{59,61}. An interesting aspect of our findings was that the most marked changes occurred in a brain region with the highest rate of protein synthesis, and presumably with one of the highest rates of ROS generation, the supraoptic nucleus of the hypothalamus⁵⁸. We also saw changes in the brainstem, hippocampus and hypothalamus as a whole⁵⁹. We hypothesize that, all other things being equal, higher metabolic rates in the brain require longer periods of sleep to interrupt ROS-induced damage to brain cells, facilitate the synthesis and activities of molecules that protect brain cells from oxidative stress, allow sufficient time for the repair or replacement of essential cellular components in neurons and glia⁶² and deal with other biochemical consequences of waking metabolic activity. This would account for an inverse relationship between body mass and sleep time.

One may hypothesize that the 'ratio' of the energy conservation

benefit of sleep to the waking metabolic-activity-derived need for sleep for brain recuperation varies across species. Carnivores and omnivores, which tend to have more sleep than predicted on the basis of body mass alone, may make more use of the energy-conservation aspects of sleep because their generally safe sleep places⁶³ and ability to eat meals with high calorific density may make continuous activity unnecessary. In such a situation, reproductive fitness might best be served by energy conservation, which would reduce the need for hunting, aid nurturing of the young and speed development. This additional sleep time would obscure any underlying sleep requirement that is linked to metabolic rate.

Protein synthesis in the brain is increased during slow-wave sleep⁶⁴. Recent work suggests another role for sleep. New neurons are generated in adult animals in the olfactory bulb, the subventricular zone lining the lateral ventricles, and in the subgranular cell layer of the dentate gyrus of the hippocampus, in a process that produces functional neurons in 3–4 weeks. This neurogenesis is facilitated by exercise and blocked by stress. Short-term (2–3-day) total sleep deprivation, even when other forms of stress are controlled for, also blocks subsequent proliferation of cells in the dentate gyrus⁶⁵. Thus, sleep may have a general role in allowing or facilitating neurogenesis.

Several substances whose activities or levels may vary with overall metabolic activity in the brain have been proposed to have roles in sleep control or function. One hypothesis was that brain glycogen is depleted during waking and restored during sleep^{66,67}; however, subsequent work showed that this effect did not occur in all strains of mice, suggesting that this could not be a general regulatory mechanism or function of sleep⁶⁸. Cytokines such as interleukin-1 have been implicated in sleep control⁶⁹. Other substances, including adenosine, prostaglandin D₂, a muramyl dipeptide, delta sleep inducing peptide, corticostatin, growth-hormone-releasing hormone, oxidized glutathione, uridine, tumour necrosis factor- α , oleamide, cortistatin, cholecystokinin, insulin, nitric oxide⁶⁹ and neuropeptide S⁷⁰, have significant effects on sleep. The relative importance of each of these 'sleep factors', the pathways by which they interact with each other and the way in which they might act on sleep-triggering mechanisms remain to be more fully elucidated.

Theories of REM sleep function

REM sleep, the state in which our most vivid dreams occur, has inspired a multitude of functional theories. The identification of the 'dream state' as a periodic physiological process during sleep⁷¹ has encouraged the addition of physiological and psychological theories to the more mystical theories of the ancients.

REM sleep is 'paradoxical' in the sense that although an animal in REM sleep is behaviourally asleep, brain metabolic and neuronal activity are high, respiration and heart rate are variable, rapid eye movements and twitches of the extremities occur and males frequently develop erections⁷². These phenomena and the vivid dreams that humans report upon awakening from REM sleep have made the function of this state particularly mysterious and intriguing. Although behavioural immobility and reduced overall body metabolic rate relative to active waking are maintained, why has this state evolved when continued NREM sleep, with its reduction in brain metabolic activity, would seem to be more efficient at achieving the recuperative and energy-saving effects of sleep?

Memory consolidation

The idea that either REM or NREM sleep is 'absolutely required' for memory consolidation has received much attention recently. I and others have reviewed this issue elsewhere and concluded that this is unlikely to be the case. Certainly disturbed sleep is not conducive to concentration and learning, but an essential role for sleep in memory consolidation remains unproven^{43,73,74}.

REM sleep and development

REM sleep amount is positively correlated with total sleep amount and

negatively correlated with body weight. However, if one statistically controls for body weight or brain weight, REM sleep amount is most strongly correlated with immaturity at birth²⁵. Altricial mammals, those that are immature at birth, tend to have more REM sleep than mammals that are mature at birth, or precocial. This tendency is marked in the neonatal period. But perhaps more remarkable is that altricial animals continue to have more REM sleep as adults. The platypus has 8 hours of REM sleep per day as an adult and the neonate cannot thermoregulate, locomote, acquire food or defend itself at birth and lives attached to its mother. The ferret, likewise, is immature at birth, and the adult has over 6 hours of REM sleep per day. By contrast, the guinea pig has only 1 hour of REM sleep per day as an adult⁷⁵. The guinea pig is born with teeth, claws, fur and open eyes; it thermoregulates at birth, locomotes within an hour of birth and eats solid food within a day of birth. Similarly, the sheep and giraffe are relatively mature at birth and have little REM sleep (less than 1 hour per day) at maturity²⁵.

Although the nature of sleep states early in the development of the rat differs from adult patterns, the brainstem mechanisms generating REM sleep are present at birth^{76,77}. The extremely high levels of REM sleep seen at birth, followed by a slow decrease to adult levels seen in many terrestrial mammals, must be an important clue to its function. This time course, combined with the observation that neuronal activity levels are high during REM sleep, led to the hypothesis that this sleep state is involved in the development of the brain. Wiesel and Hubel discovered that reducing light input into one eye for a period of several days in neonates led to a reduction in the number of cells in the visual system receiving input from that eye. This reduction persisted into adulthood, long after normal visual input had been restored. Monocular light deprivation during the 'critical' neonatal period also caused a reduction in the thickness of the lateral geniculate layers innervated by the closed eye. Others subsequently found that when animals were also REM sleep deprived during the critical period of susceptibility, this shrinkage was accelerated⁷⁸. This finding indicates that the activity of the visual system known to occur in REM sleep normally compensates for any asymmetrical, abnormal or absent input, preventing processes that prune away unused connections during development. One can imagine that REM sleep serves this function for other sensory systems and perhaps for motor systems as well, given the intense central motor activation that occurs during this state (whose peripheral expression is blocked by the inhibition of motoneurons)^{14,19}. As long as the extended NREM sleep of neonates is interrupted by the increased neuronal activity of REM sleep, neuronal development proceeds according to genetic programmes⁷⁹.

REM sleep in adults

If we accept that the large amount of REM sleep early in life serves to maintain or establish brain connections during crucial periods of development, what is the function of this state later in life? One idea that has been proposed repeatedly is that REM sleep stimulates the adult brain during sleep to reverse the effects of NREM sleep on immediately subsequent waking behaviour. Animals typically awaken spontaneously from REM sleep⁸⁰. Animals that are awakened from NREM sleep have poor sensory-motor function compared with those awakened from REM sleep⁸¹. Awakening in a more alert state would convey a substantial selective advantage.

In humans, the duration of REM sleep episodes progressively increases throughout the sleep period and is maximal at the expected time of awakening. The initial REM sleep period of the night may last only 5–10 minutes, whereas the last REM sleep period may last more than 25 minutes. The intensity of REM sleep, as measured by the density of eye movements and twitches, the prevalence of erections in males and the vividness of dream reports also increases as the night progresses. REM sleep amounts are maximal near the nadir of the brain and core body temperature cycles. And temperature in several brain regions increases during REM sleep^{82,83}, even though the regulation of body temperature is largely suppressed during REM sleep¹⁵.

One would not expect an increase in duration and intensity of REM sleep across the sleep period if REM sleep intensity were linked to one or more aspects of the previous waking episode. If a strong relationship with previous waking existed, one would expect to see maximal REM sleep intensity and duration in the early part of the night, alternating with NREM sleep, which is most intense at this time. When ambient temperature is outside the thermoneutral range, REM sleep is suppressed more strongly than is NREM sleep¹⁵, suggesting that the preservation of NREM sleep is of higher priority than that of REM sleep.

The absence or very small amounts of REM sleep in marine mammals displaying unihemispheric sleep with brainstem-regulated motor activity supports the hypothesis that the stimulation of brainstem-activating systems is an important function of REM sleep. In these animals, the presence of continuous motor activity would be expected to maintain a high level of metabolic activity in brainstem structures 24 hours a day, eliminating any need for REM sleep activation of these systems. Similarly, the manifestation of REM sleep in monotremes as a largely brainstem state, without marked neocortical activation, suggests that REM sleep may have evolved as a state of brainstem activation, with cortical stimulation functions added later in evolution. The cold-induced increase in REM sleep amount in the isolated brainstem¹⁷, the increased REM sleep amount at the minimum of the circadian brain and body temperature cycles and the increase in the temperatures of brain regions during REM sleep^{82,83} are consistent with this brainstem-activation hypothesis.

Surprisingly, in humans, one can increase the amount of REM sleep and percentage of total sleep time devoted to this state simply by extending sleep time. Furthermore, this increased REM sleep time does not seem to produce a homeostatic reduction in REM sleep time during the subsequent night^{42,84}. Although anxiety can decrease the amount of REM sleep⁸⁵, acute stress can greatly increase REM sleep time in the rat, despite the absence of sleep loss during the previous stress⁸⁶, suggesting that under some conditions REM sleep 'rebound' may be related to the emotional or autonomic concomitants of the REM sleep-deprivation procedure rather than the loss of REM sleep itself.

The proposed functions of REM sleep in adults are consistent with the lack of any easily detectable cognitive or physiological symptoms in humans in whom REM sleep has been suppressed for months or years by monoamine oxidase inhibitors or brain lesions⁴³. One might hypothesize that these individuals would be less alert upon waking, a change that would have minimal survival implications in most contemporary humans but might have significant effects upon reproductive success in other animals.

Regulation of monoaminergic systems

The linked cessation of activity in noradrenaline, serotonin, histamine and hypocretin cells that occurs during normal REM sleep, and to a lesser extent during NREM sleep, suggests another possible function for sleep. Sleep may resensitize these REM off or sleep off transmitter systems, which are tonically active during waking, by increasing the quantities and activities of their synthetic enzymes, transporter mechanisms and receptors^{87,88}. Three studies favour a role for REM sleep in the maintenance of central noradrenergic receptors, whereas one study questions this role. If this hypothesis were valid, it seems clear that the nature of such resensitization would be specific for particular receptor types and brain regions^{89–92}.

REM sleep deprivation has antidepressive effects. It may be that a normal function of REM sleep is to dampen activity and emotional expression^{42,93} by causing the changes in monoaminergic systems proposed above. These systems are well known to be involved in emotional regulation, and most antidepressive medications act through effects on monoaminergic systems and suppress REM sleep.

REM sleep deprivation causes an acceleration of body heat loss during subsequent waking in the rat^{94,95}. REM sleep itself is a state in which thermoregulation is attenuated⁹⁶, suggesting that recuperative changes

in thermoregulatory control mechanisms may occur during REM sleep periods, perhaps through the maintenance, repair or sensitization of peripheral monoaminergic control mechanisms, in much the same way that an improvement in the central functioning of these systems has been proposed to take place during REM sleep. The unique thermoregulatory challenges faced by cetaceans may also be a factor in the apparent absence of significant amounts of REM sleep in this order. Although a high level of heat generation is vital to maintaining their body temperature, sea temperature varies little over the 24-hour day, making the 'recalibration' proposed for thermoregulatory systems during their relative suspension in REM sleep less necessary⁸.

Future directions

Sleep probably has multiple functions for the brain and body. An important task will be the identification of which of the hypothesized functions may only be achieved during sleep, and which may be executed during both waking and sleep, with sleep being a more efficient time for their accomplishment. It also remains to be determined which, if any, of the proposed functions are universal across mammalian species and across the lifespan, and which may be limited to particular species or phases of development.

- Dinges, D. F., Rogers, N. L. & Baynard, M. D. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 67–76 (Elsevier Saunders, Philadelphia, 2005).
- Huber, R. *et al.* Sleep homeostasis in *Drosophila melanogaster*. *Sleep* **27**, 628–639 (2004).
- Czeisler, C. A., Buxton, O. & Khalsa, S. B. S. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 375–394 (Elsevier Saunders, Philadelphia, 2005).
- Bell-Pedersen, D. *et al.* Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Rev. Genet.* **6**, 554–556 (2005).
- Everson, C. A., Smith, C. B. & Sokoloff, L. Effects of prolonged sleep deprivation on local rates of cerebral energy metabolism in freely moving rats. *J. Neurosci.* **14**, 6769–6778 (1994).
- Nofzinger, E. A. *et al.* Functional neuroimaging evidence for hyperarousal in insomnia. *Am. J. Psychiatry* **161**, 2126–2128 (2004).
- Maquet, P. *et al.* Regional organisation of brain activity during paradoxical sleep (PS). *Arch. Ital. Biol.* **142**, 413–419 (2004).
- Rechtschaffen, A. Current perspectives on the function of sleep. *Perspect. Biol. Med.* **41**, 359–390 (1998).
- Cirelli, C. *et al.* Reduced sleep in *Drosophila Shaker* mutants. *Nature* **434**, 1087–1092 (2005).
- Kume, K., Kume, S., Park, S. K., Hirsh, J. & Jackson, F. R. Dopamine is a regulator of arousal in the fruit fly. *J. Neurosci.* **25**, 7377–7384 (2005).
- Kryger, M. H., Roth, T. & Dement, W. C. *Principles and Practice of Sleep Medicine* (Elsevier Saunders, Philadelphia, 2005).
- McGinty, D. J. & Szymusiak, R. S. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 169–184 (Elsevier Saunders, Philadelphia, 2005).
- Saper, C. B., Chou, T. C. & Scammell, T. E. The sleep switch: hypothalamic control of sleep and wakefulness. *Trends Neurosci.* **24**, 726–731 (2001).
- Siegel, J. M. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 120–135 (Elsevier Saunders, Philadelphia, 2005).
- Heller, H. C. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 292–304 (Elsevier Saunders, Philadelphia, 2005).
- Coleman, C. G., Lydic, R. & Baghdoyan, H. A. M2 muscarinic receptors in pontine reticular formation of C57BL/6J mouse contribute to rapid eye movement sleep generation. *Neuroscience* **126**, 821–830 (2004).
- Jouvet, M., Buda, C., Debilly, G., Dittmar, A. & Sastre, J. P. Glycine immunoreactive neurons in the medulla oblongata in cat. *C. R. Acad. Sci. III* **306**, 69–73 (1988).
- Steriade, M. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 101–119 (Elsevier Saunders, Philadelphia, 2005).
- Chase, M. H. & Morales, F. R. in *Principles of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 154–168 (Elsevier Saunders, Philadelphia, 2005).
- Aston-Jones, G. & Bloom, F. E. Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep-waking cycle. *J. Neurosci.* **1**, 876–886 (1981).
- Jacobs, B. L. & Azmitia, E. C. Structure and function of the brain serotonin system. *Physiol. Rev.* **72**, 165–229 (1992).
- McGinty, D. J. & Harper, R. M. Dorsal raphe neurons: depression of firing during sleep in cats. *Brain Res.* **101**, 569–575 (1976).
- Mileykovskiy, B. Y., Kiyashchenko, L. I. & Siegel, J. M. Behavioral correlates of activity in identified hypocretin/orexin neurons. *Neuron* **46**, 787–798 (2005).
- John, J., Wu, M.-F., Boehmer, L. N. & Siegel, J. M. Cataplexy-active neurons in the hypothalamus: implications for the role of histamine in sleep and waking behavior. *Neuron* **42**, 619–634 (2004).
- Zepelin, H., Siegel, J. M. & Tobler, I. in *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 91–100 (Elsevier Saunders, Philadelphia, 2005).
- Grutzner, F. *et al.* In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432**, 913–917 (2004).
- Allison, T. & Van Twyver, H. Electrophysiological studies of the echidna, *Tachyglossus aculeatus*. II. Dormancy and hibernation. *Arch. Ital. Biol.* **110**, 145–184 (1972).
- Siegel, J. M., Manger, P., Nienhuis, R., Fahringer, H. M. & Pettigrew, J. The echidna

- Tachyglossus aculeatus* combines REM and non-REM aspects in a single sleep state: implications for the evolution of sleep. *J. Neurosci.* **16**, 3500–3506 (1996).
29. Siegel, J. M. et al. Sleep in the platypus. *Neuroscience* **91**, 391–400 (1999).
 30. Mukhametov, L. M., Supin, A. Y. & Polyakova, I. G. Interhemispheric asymmetry of the electroencephalographic sleep patterns in dolphins. *Brain Res.* **134**, 581–584 (1977).
 31. Lyamin, O. I., Mukhametov, L. M. & Siegel, J. M. Relationship between sleep and eye state in Cetaceans and Pinnipeds. *Arch. Ital. Biol.* **142**, 557–568 (2004).
 32. Mukhametov, L. M., Lyamin, O. I., Chetyrbok, I. S., Vassilyev, A. A. & Diaz, R. P. Sleep in an Amazonian manatee, *Trichechus inunguis*. *Experientia* **48**, 417–419 (1992).
 33. Mukhametov, L. M., Lyamin, O. I., Shpak, O. V., Manger, P. & Siegel, J. M. Swimming styles and their relationship to rest and activity states in captive Commerson's dolphins. Proc. 14th Biennial Conference on the Biology of Marine Mammals 152 (2002).
 34. Vanderwolf, C. H. & Baker, G. B. Evidence that serotonin mediates non-cholinergic neocortical low voltage fast activity, non-cholinergic hippocampal rhythmic slow activity and contributes to intelligent behavior. *Brain Res.* **374**, 342–356 (1986).
 35. Bergmann, B. M., Winter, J. B., Rosenberg, R. S. & Rechtschaffen, A. NREM sleep with low-voltage EEG in the rat. *Sleep* **10**, 1–11 (1987).
 36. Oleksenko, A. I., Mukhametov, L. M., Polyakova, I. G., Supin, A. Y. & Kovalzon, V. M. Unihemispheric sleep deprivation in bottlenose dolphins. *J. Sleep Res.* **1**, 40–44 (1992).
 37. Siegel, J. M. & Tomaszewski, K. S. Behavioral organization of reticular formation: studies in the unrestrained cat. I. Cells related to axial, limb, eye, and other movements. *J. Neurophysiol.* **50**, 696–716 (1983).
 38. Lyamin, O., Pryaslova, J., Lance, V. & Siegel, J. Animal behaviour: continuous activity in cetaceans after birth. *Nature* **435**, 1177 (2005).
 39. Carskadon, M. A. & Dement, W. C. In *Principles and Practice of Sleep Medicine* Vol. 4 (eds Kryger, M. H., Roth, T. & Dement, W. C.) 13–23 (Elsevier Saunders, Philadelphia, 2005).
 40. Rattenborg, N. C. et al. Migratory sleeplessness in the white-crowned sparrow (*Zonotrichia leucophrys gambelii*). *PLoS Biol.* **2**, E212 (2004).
 41. Lyamin, O. I., Oleksenko, A. I., Polyakova, I. G. & Mukhametov, L. M. Paradoxical sleep in northern fur seals in water and on land. *J. Sleep Res.* **5** (suppl.), 130–130 (1996).
 42. Horne, J. A. REM sleep — by default? *Neurosci. Biobehav. Rev.* **24**, 777–797 (2000).
 43. Siegel, J. M. The REM sleep-memory consolidation hypothesis. *Science* **294**, 1058–1063 (2001).
 44. Huber, R., Ghilardi, M. F., Massimini, M. & Tononi, G. Local sleep and learning. *Nature* **430**, 78–81 (2004).
 45. Krueger, J. M., Obal, F. J. & Fang, J. Why we sleep: a theoretical view of sleep function. *Sleep Med. Rev.* **3**, 119–129 (1999).
 46. Vyazovskiy, V. V., Welker, E., Fritschy, J. M. & Tobler, I. Regional pattern of metabolic activation is reflected in the sleep EEG after sleep deprivation combined with unilateral whisker stimulation in mice. *Eur. J. Neurosci.* **20**, 1363–1370 (2004).
 47. Achermann, P. & Borbély, A. A. Mathematical models of sleep regulation. *Front. Biosci.* **8**, s683–s693 (2003).
 48. Verret, L., Leger, L., Fort, P. & Luppi, P. H. Cholinergic and noncholinergic brainstem neurons expressing Fos after paradoxical (REM) sleep deprivation and recovery. *Eur. J. Neurosci.* **21**, 2488–2504 (2005).
 49. Binks, P. G., Waters, W. F. & Hurry, M. Short-term total sleep deprivation does not selectively impair higher cortical functioning. *Sleep* **22**, 328–334 (1999).
 50. Balkin, T. J. et al. On the importance of countermeasures in sleep and performance models. *Aviat. Space Environ. Med.* **75**, A155–A157 (2004).
 51. Villablanca, J. R. Counterpointing the functional role of the forebrain and of the brainstem in the control of the sleep-waking system. *J. Sleep Res.* **13**, 179–208 (2004).
 52. Everson, C. A. & Crowley, W. R. Reductions in circulating anabolic hormones induced by sustained sleep deprivation in rats. *Am. J. Physiol. Endocrinol. Metab.* **286**, E1060–E1070 (2004).
 53. O'Hara, B. F. et al. Gene expression in the brain across the hibernation cycle. *J. Neurosci.* **19**, 3781–3790 (1999).
 54. Lima, S. L., Rattenborg, N. C., Lesku, J. A. & Amlaner, C. J. Sleep under the risk of predation. *Anim. Behav.* **70**, 723–736 (2005).
 55. Merrick, A. W. & Scharp, D. W. Electroencephalography of resting behavior in cattle, with observations on the question of sleep. *Am. J. Vet. Res.* **32**, 1893–1897 (1971).
 56. Klemm, W. R. Sleep and paradoxical sleep in ruminants. *Proc. Soc. Exp. Biol. Med.* **121**, 635–638 (1966).
 57. Turner, N., Else, P. L. & Hulbert, A. J. An allometric comparison of microsomal membrane lipid composition and sodium pump molecular activity in the brain of mammals and birds. *J. Exp. Biol.* **208**, 371–381 (2005).
 58. Eiland, M. M. et al. Increases in amino-cupric-silver staining of the supraoptic nucleus after sleep deprivation. *Brain Res.* **945**, 1–8 (2002).
 59. Ramanathan, L., Gulyani, S., Nienhuis, R. & Siegel, J. M. Sleep deprivation decreases superoxide dismutase activity in rat hippocampus and brainstem. *Neuroreport* **13**, 1387–1390 (2002).
 60. Everson, C. A., Laatsch, C. D. & Hogg, N. Antioxidant defense responses to sleep loss and sleep recovery. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **288**, R374–R383 (2005).
 61. Gopalakrishnan, A., Ji, L. L. & Cirelli, C. Sleep deprivation and cellular responses to oxidative stress. *Sleep* **27**, 27–35 (2004).
 62. Zimmerman, J. E., Mackiewicz, M., Galante, R. J. et al. Glycogen in the brain of *Drosophila melanogaster*: diurnal rhythm and the effect of rest deprivation. *J. Neurochem.* **88**, 32–40 (2004).
 63. Allison, T. & Cicchetti, D. V. Sleep in mammals: ecological and constitutional correlates. *Science* **194**, 732–734 (1976).
 64. Nakanishi, H. et al. Positive correlations between cerebral protein synthesis rates and deep sleep in Macaca mulatta. *Eur. J. Neurosci.* **9**, 271–279 (1997).
 65. Guzman-Marín, R. et al. Sleep deprivation reduces proliferation of cells in the dentate gyrus of the hippocampus in rats. *J. Physiol.* **549**, 563–571 (2003).
 66. Benington, J. H. & Heller, H. C. Restoration of brain energy metabolism as the function of sleep. *Prog. Neurobiol.* **45**, 347–360 (1995).
 67. Kong, J. et al. Brain glycogen decreases with increased periods of wakefulness: implications for homeostatic drive to sleep. *J. Neurosci.* **22**, 5581–5587 (2002).
 68. Franken, P., Gip, P., Hagiwara, G., Ruby, N. F. & Heller, H. C. Changes in brain glycogen after sleep deprivation vary with genotype. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **285**, R413–R419 (2003).
 69. Opp, M. R. & Krueger, J. M. Interleukin-1 is involved in responses to sleep deprivation in the rabbit. *Brain Res.* **639**, 57–65 (1994).
 70. Xu, Y. L. et al. Neuropeptide S: a neuropeptide promoting arousal and anxiolytic-like effects. *Neuron* **43**, 487–497 (2004).
 71. Aserinsky, E. & Kleitman, N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science* **118**, 273–274 (1953).
 72. Schmidt, M. H., Valatx, J. L., Sakai, K., Fort, P. & Jouvet, M. Role of the lateral preoptic area in sleep-related erectile mechanisms and sleep generation in the rat. *J. Neurosci.* **20**, 6640–6647 (2000).
 73. Siegel, J. M. & Vertes, R. P. Sleep and memory: The ongoing debate, Rebuttal. *Sleep* **28**, 1232–1233 (2005).
 74. Vertes, R. P. Memory consolidation in sleep; dream or reality. *Neuron* **44**, 135–148 (2004).
 75. Jouvet-Mounier, D., Astic, L. & Lacote, D. Ontogenesis of the states of sleep in rat, cat, and guinea pig during the first postnatal month. *Dev. Psychobiol.* **2**, 216–239 (1970).
 76. Frank, M. G. & Heller, H. C. The ontogeny of mammalian sleep: a reappraisal of alternative hypotheses. *J. Sleep Res.* **12**, 25–34 (2003).
 77. Karlsson, K. A. E., Gall, A. J., Mohs, E. J., Seelke, A. M. H. & Blumberg, M. S. The neural substrates of infant sleep in rats. *PLoS Biol.* **3**, e143 (2005).
 78. Shaffery, J. P., Roffwarg, H. P., Speciale, S. G. & Marks, G. A. Ponto-geniculate-occipital-wave suppression amplifies lateral geniculate nucleus cell-size changes in monocularly deprived kittens. *Brain Res. Dev. Brain Res.* **114**, 109–119 (1999).
 79. Jouvet, M. In *Cerebral Correlates of Conscious Experience*. INSERM Symposium Vol. 6 (eds Buser, P. & Rougeul-Buser, A.) 245–261 (Elsevier/North-Holland Biomed., Amsterdam, 1978).
 80. Snyder, F. Toward an evolutionary theory of dreaming. *Am. J. Psychiatry* **123**, 121–136 (1966).
 81. Horner, R. L., Sanford, L. D., Pack, A. I. & Morrison, A. R. Activation of a distinct arousal state immediately after spontaneous awakening from sleep. *Brain Res.* **778**, 127–134 (1997).
 82. Baker, F. C., Angara, C., Szymusiak, R. & McGinty, D. Persistence of sleep-temperature coupling after suprachiasmatic nuclei lesions in rats. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* (2005).
 83. Wehr, T. A. A brain-warming function for REM sleep. *Neurosci. Biobehav. Rev.* **16**, 379–397 (1992).
 84. Darchia, N., Campbell, I. G., Palagini, L. & Feinberg, I. Rapid eye movement density shows trends across REM periods but is uncorrelated with NREM delta in young and elderly human subjects. *Brain Res. Bull.* **63**, 433–438 (2004).
 85. Sanford, L. D., Tang, X., Ross, R. J. & Morrison, A. R. Influence of shock training and explicit fear-conditioned cues on sleep architecture in mice: strain comparison. *Behav. Genet.* **33**, 43–58 (2003).
 86. Gonzalez, M. M., Debilly, G., Valatx, J. L. & Jouvet, M. Sleep increase after immobilization stress: role of the noradrenergic locus coeruleus system in the rat. *Neurosci. Lett.* **202**, 5–8 (1995).
 87. Shaw, P. J., Cirelli, C., Greenspan, R. J. & Tononi, G. Correlates of sleep and waking in *Drosophila melanogaster*. *Science* **287**, 1834–1837 (2000).
 88. Siegel, J. M. & Rogawski, M. A. A function for REM sleep: regulation of noradrenergic receptor sensitivity. *Brain Res. Rev.* **13**, 213–233 (1988).
 89. Pedrazzoli, M. & Benedito, M. A. Rapid eye movement sleep deprivation-induced down-regulation of beta-adrenergic receptors in the rat brainstem and hippocampus. *Pharmacol. Biochem. Behav.* **79**, 31–36 (2004).
 90. Tsai, L. L., Bergmann, B., Perry, B. & Rechtschaffen, A. Effects of chronic total sleep deprivation on central noradrenergic receptors in rat brain. *Brain Res.* **602**, 221–227 (1993).
 91. Hipolide, D. C., Tufik, S., Raymond, R. & Nobrega, J. N. Heterogeneous effects of rapid eye movement sleep deprivation on binding to alpha- and beta-adrenergic receptor subtypes in rat brain. *Neuroscience* **86**, 977–987 (1998).
 92. Hipolide, D. C. et al. Distinct effects of sleep deprivation on binding to norepinephrine and serotonin transporters in rat brain. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **29**, 297–303 (2005).
 93. Vogel, G. W. An alternative view of the neurobiology of dreaming. *Am. J. Psychiatry* **135**, 1531–1535 (1978).
 94. Rechtschaffen, A. & Bergmann, B. M. Sleep deprivation in the rat: an update of the 1989 paper. *Sleep* **25**, 18–24 (2002).
 95. Zenko, C. E., Bergmann, B. M. & Rechtschaffen, A. Vascular resistance in the rat during baseline, chronic total sleep deprivation, and recovery from total sleep deprivation. *Sleep* **23**, 341–346 (2000).
 96. Parmeggiani, P. L., Azzaroni, A. & Calasso, M. Systemic hemodynamic changes raising brain temperature in REM sleep. *Brain Res.* **940**, 55–60 (2002).
 97. Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates* (Elsevier Academic Press, London, 2005).
 98. Lyamin, O. I. et al. Unihemispheric slow wave sleep and the state of the eyes in a white whale. *Behav. Brain Res.* **129**, 125–129 (2002).
 99. Mukhametov, L. M. & Polyakova, I. G. [Electroencephalographic study of sleep in Sea of Azov porpoises.] *Zh. Vyssh. Nerv. Deiat. Im. I. P. Pavlova* **31**, 333–339 (1981).
 100. Mukhametov, L. M. Unihemispheric slow-wave sleep in the Amazonian dolphin, *Inia geoffrensis*. *Neurosci. Lett.* **79**, 128–132 (1987).

Acknowledgements Supported by NIH and NSF and DARPA. I thank O. Lyamin for the beluga and dolphin photo and the graph of beluga sleep, and A. Siegel, L. Boehmer, R. Nienhuis, A. Rechtschaffen, I. Tobler, C. Heller, S. Ridgway, J. Horne, D. McGinty, C. Amlaner and J. Lesku for very helpful comments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing interests. Correspondence should be addressed to J.M.S. (JSiegel@ucla.edu).

Sleep-dependent memory consolidation

Robert Stickgold¹

The concept of ‘sleeping on a problem’ is familiar to most of us. But with myriad stages of sleep, forms of memory and processes of memory encoding and consolidation, sorting out how sleep contributes to memory has been anything but straightforward. Nevertheless, converging evidence, from the molecular to the phenomenological, leaves little doubt that offline memory reprocessing during sleep is an important component of how our memories are formed and ultimately shaped.

The question of how sleep might contribute to learning and memory consolidation is an old one. In the first century AD, the Roman rhetorician Quintilian, commenting on the benefits of sleep, noted that, “what could not be repeated at first is readily put together on the following day; and the very time which is generally thought to cause forgetfulness is found to strengthen the memory”¹. Although this may have been obvious to him, it has been less obvious to the research community, and, until a seminal paper by Karni, Sagi and colleagues in 1994 (ref. 2), the topic received relatively little attention within either the sleep or memory research communities. But over the past 10 years, the rate of publication of research papers on sleep-dependent learning and memory consolidation has increased fivefold³. Evidence supporting sleep-dependent memory consolidation has come from a range of molecular, cellular, physiological and behavioural studies (for a review, see ref. 4; for an opposing view, see ref. 5).

One of the major problems facing this area of research is that the terms sleep, memory and memory consolidation all refer to complex phenomena, none of which can be treated as a singular event. I begin this review by clarifying my use of these terms, and then present some of the more convincing evidence from studies of procedural learning in humans. I then review more broadly the behavioural evidence for sleep-dependent consolidation of perceptual and motor skill procedural memories, declarative memories and complex cognitive procedural memories. I follow this by outlining converging evidence from molecular, cellular, neurophysiological, brain-imaging and dream studies, all of which support an important, and sometimes essential, role for sleep in memory consolidation.

Sleep and memory

There is more than one type of memory. Consider, for example, the capital of France, what you had for dinner last night and how to ride a bicycle. All three of these recollections require information that you have learned and stored, but they are very different types of memory. Multiple memory systems store different classes of memory in different brain regions and, quite probably, in different formats.

Memories are most commonly divided into declarative memories, which are those that a person can call to mind (for example, the capital of France or last night's dinner), and non-declarative memories, which are those that are normally used without conscious recollection (for example, how to ride a bicycle or how to talk your way out of a parking ticket) (Fig. 1)⁶. Declarative memories are further divided into episodic memories, that is, memories of specific events (such as what you had for dinner last night), and semantic memories, in other words memories of general information (such as the capital of France). Non-

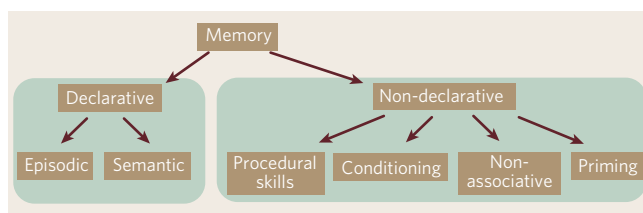


Figure 1 | Categorization of memory systems.

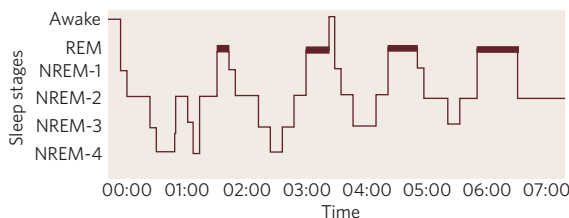


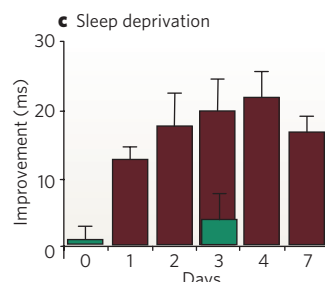
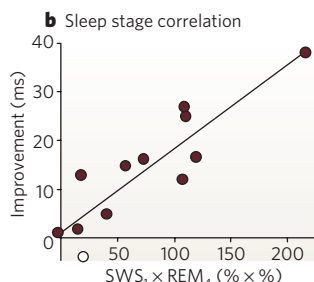
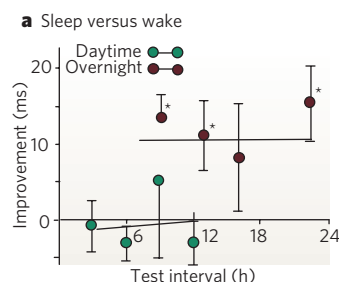
Figure 2 | Categorization of sleep stages.

declarative memories are also divided into several subcategories, such as procedural skills. Others in the field use the term memory in a more restricted manner, such as limiting it to episodic memories (see the commentary by Hobson in this issue, p. 1254).

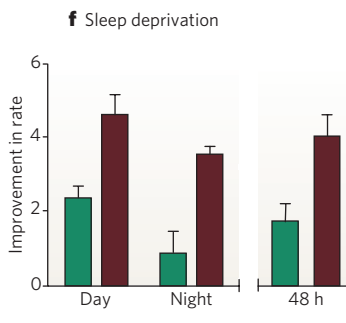
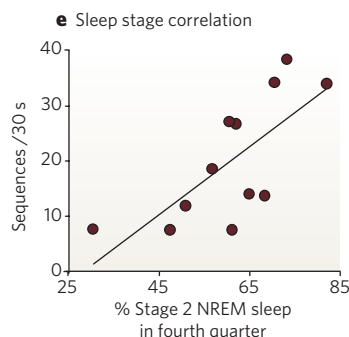
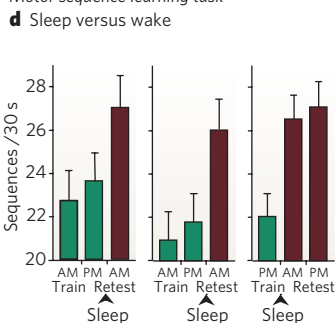
There is no consensus on what processes should be covered by the term ‘memory consolidation’. The term memory consolidation originally referred to a process of memory stabilization, through which memories become resistant to interference^{7,8}. But after the initial encoding of a sensorimotor experience, a series of cellular, molecular and systems-level alterations develop over time, automatically and outside of awareness, that stabilize and enhance the initial memory representation, converting it into a long-lasting and optimally integrated memory. These include not only cellular and molecular processes occurring at the local synaptic level, but systems-level reorganizations of individual memories as well (see refs 9, 10, for example). These additional memory-consolidation processes show the greatest evidence of sleep dependence, and I include all of them under the umbrella term memory consolidation. It is important to note that there is no consensus on how many distinct post-encoding processes exist, how they should be defined, and which should be considered under the rubric of memory consolidation. For example, memory stabilization cannot be considered absolute, because that would mean that other consolidation processes, such as enhancement and reorganization of memories, would not be possible.

¹Department of Psychiatry, Harvard Medical School, and Centre for Sleep and Cognition, Beth Israel Deaconess Medical Centre, 330 Brookline Avenue/FD-861, Boston, Massachusetts 02215, USA.

Visual texture discrimination task



Motor sequence learning task



Motor adaption learning task

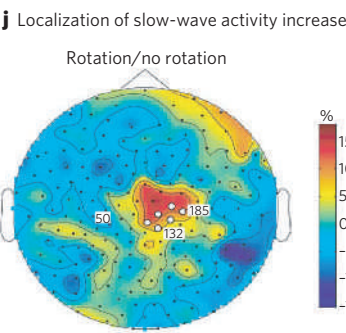
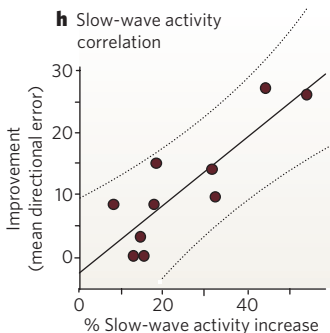
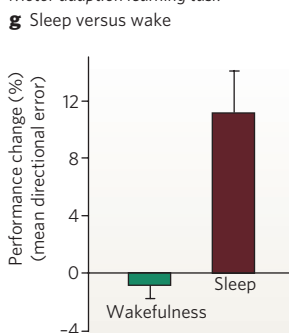


Figure 3 | Sleep-dependent consolidation of procedural memories. **a–c**, Participants in a visual texture discrimination task show improvement only after post-training sleep, even when finger movement was suppressed with mittens during wake periods (**d**). Improvement correlated with early-night slow-wave sleep and late-night REM sleep. **d–f**, Participants in a motor sequence finger-tapping task show similar sleep-dependent improvement, correlated with late-night stage 2 non-REM sleep. **g–j**, Participants in a motor adaption task also show sleep-dependent improvement, correlated with EEG slow-wave activity in task-related regions of the cortex. All error bars represent s.e.m. See the text for further details. Green bars, performance without intervening sleep or without sleep on the first post-training night. Dark red bars, performance after normal sleep. See the text for further details. Panel J is reproduced with permission from ref. 21.

When we speak of consolidation processes as being 'sleep dependent', we are hypothesizing that they occur primarily during sleep. But most sleep-related processes can occur during periods of wakefulness and vice versa (for example, waking hallucinations and sleep walking). Nonetheless, in most cases, experimental studies have found sleep-dependent processes occurring only during sleep.

Unlike the nature of memory and memory consolidation, the structure of human sleep is clear. A night of sleep is composed of ~90-minute cycles divided into periods of rapid eye-movement sleep (REM) and non-REM sleep (NREM), with NREM further divided into stages 1 to 4 (Fig. 2). Stages 3 and 4 are the deepest stages of sleep, and are referred to collectively as slow-wave sleep (SWS) on the basis of the patterns of large, slow (0.5–4-Hz) oscillations observed on the electroencephalogram (EEG). Sleep stages differ not only in depth of sleep but also in the frequency and intensity of dreaming, EEG oscillations, eye movements, muscle tone, neuromodulation of cortical circuits, regional brain activation and communication between memory systems¹¹. REM, stage 2 and SWS have all been implicated in sleep-dependent memory processing, as have specific patterns of synchronous cortical neuronal activation associated with these stages, including ponto-occipitogeniculate waves and theta rhythms in REM, sleep spindles in stage 2 and the slow-wave activity of SWS.

Sleep-dependent memory enhancement

Results from three laboratories who asked volunteers to perform three different tasks, a visual texture discrimination test^{12,13}, a motor

sequence test^{14,15} and a motor adaption test¹⁶, demonstrated that all subjects show post-training improvement after a night's sleep but not during an equivalent period of being awake. The results also demonstrated that the amount of overnight improvement correlates with the amount of specific sleep stages or sleep events, and that total or partial sleep deprivation can prevent the normal overnight improvement.

These results leave little doubt that sleep contributes to the consolidation of memories, especially their enhancement. But the details of these processes remain unclear. Although improvement on the visual texture discrimination task correlates with the levels of REM and SWS, improvements in the motor sequence task correlate with lighter stage 2 NREM. With the motor adaption task, improvement was linked, using EEG patterns, to SWS. Thus, even for these procedural skill tasks, sleep-dependent consolidation does not seem to depend consistently on one specific aspect of sleep. Instead, each stage of sleep seems to contribute differently to these processes, and we have proposed that the multiplicity of sleep stages has evolved, in part, to provide optimal brain states for a range of distinct memory consolidation processes. I will now look at each of the tests in more detail.

Visual texture discrimination

Several studies have supported the sleep-dependent enhancement of a foreground-background discrimination task (Fig. 3a–c). In this task, subjects must identify the orientation of an array of diagonal bars against a background grid of horizontal bars. Although an initial study found overnight improvement blocked by REM deprivation, but not

by SWS deprivation², subsequent studies suggested roles for both SWS and REM^{12,17}. For example, overnight improvement is strongly correlated with both the amount of SWS in the first quarter of the night ($r = 0.70$, $P = 0.01$) and the amount of REM in the last quarter ($r = 0.76$, $P = 0.004$), leading to the suggestion that a two-step process is necessary, the first during SWS and the second during REM¹². The product of these two sleep parameters, $\text{SWS1} \times \text{REM4}$, produces an even stronger correlation ($r = 0.89$, $P < 0.0001$; Fig. 3b)¹². In addition, sleep deprivation the night after training prevents normal improvement, even when measured 72 hours after training, following two nights of recovery sleep (Fig. 3c, 3 days, green bar)¹³. Significant improvement is also seen after the first 3 hours of the night, rich in SWS (means: SWS = 74 min, REM = 24 min) but not across the last 3 hours of REM-rich sleep (SWS = 32 min, REM = 58 min)¹⁷. But an entire night, including both SWS-rich and REM-rich periods, shows more than three times as much improvement as the early night period¹⁷.

After consolidation takes place during this first post-training night, no improvement is seen throughout the second day (Fig. 3a). But additional nights of sleep produce further improvements (Fig. 3c). Thus, the sleep-dependent process of memory enhancement continues for at least 48–96 hours, an order of magnitude longer than the time thought to be required for the stabilization component of consolidation.

Motor sequence task

In this finger-tapping task, subjects type a simple numeric sequence, such as 4–1–3–2–4, on a computer keyboard, as quickly and accurately as possible. Alternative versions require subjects to touch fingers to thumb in a particular order. In the keyboard version, subjects improve steadily over the first minutes of practice, and then approach an asymptote, with a ~60% improvement in speed over twelve 30-second trials¹⁴. Subsequent retesting, 4–12 hours later the same day, shows no significant improvement (4%, $P = 0.13$; Fig. 3d, left panel, green bars). In contrast, after a night of sleep a significant 20% increase in speed is seen ($P > 0.0001$; Fig. 3d, left panel, red bar)¹⁴, an increase that correlates with the amount of stage 2 NREM in the last quarter of the night ($r = 0.72$, $P = 0.008$; Fig. 3e). Although daytime wake produces no improvement, a 90-minute midday nap leads to a significant 16% improvement¹⁸. Retesting after 72 hours shows even more improvement than after 24 hours (26% compared with 17%; $P = 0.07$), suggesting that, as with the visual discrimination task, improvement continues beyond the first night¹⁹. With the finger-thumb opposition version, sleep deprivation the night after training reduces the amount of improvement by over 75% ($P < 0.001$) after one night of recovery sleep (Fig. 3f, right panel). Again, sleep during the day produces significantly more improvement than an equivalent period of daytime wake ($P < 0.001$; Fig. 3f, left panel)¹³.

In addition to improvements in speed, sleep also leads to an increase in accuracy. Although subjects show no significant improvement in either speed or accuracy during 12 hours of daytime wake (speed: +4%, $P = 0.13$; error rate: +9%, $P = 0.46$), they improve significantly across a night of sleep (speed: +20%, $P < 0.0001$; error rate: -36%, $P = 0.01$)¹³.

Compared with memory enhancement, the issue of memory stabilization is more complex. Cross-training on a second finger-tapping sequence 10 minutes after training does not affect the speed or accuracy of performance on the initial sequence when retested shortly thereafter²⁰. From this perspective, consolidation seems complete within 10 minutes of training.

But when retesting is performed the next day, a different picture is seen. Now, the prior cross-training is seen to interfere with normal overnight enhancement of accuracy. Although both sequences show normal overnight improvement in speed the next day (sequence 1 = 12%, $P < 0.005$; sequence 2 = 6%, $P < 0.001$), the original sequence no longer shows the normal decrease in error rate (sequence 1 = -1%, $P = 0.90$; sequence 2 = -39%, $P < 0.001$). Thus, cross-training seems to

interfere not with the subject's recently gained ability to perform the task, but rather with the subsequent sleep-dependent enhancement of that ability. This susceptibility to interference seems to be time limited, because cross training 6 hours after the initial sequence fails to block overnight improvement in either speed or accuracy. In this case, both sequences show overnight improvement in accuracy (sequence 1 = 20%, $P < 0.05$; sequence 2 = 37%, $P < 0.005$). Whether stabilization is complete at 6 hours is unclear; the first sequence showed only half (54%) of the increase in accuracy seen in controls, but the difference was not statistically significant²⁰. But in either case, this stabilization now occurs over a period of wakefulness rather than sleep. When sleep can similarly support this stabilization is unknown.

Thus, this procedural motor sequence task shows features of sleep-dependent enhancement similar to those reported for the visual texture discrimination task. Performance improves significantly over a night of sleep, but not over either an equivalent period of wake during the day or a night of sleep deprivation; overnight improvement correlates with amounts of one or more specific (but different) sleep stages; and performance improves with a daytime nap. In contrast, memory stabilization can occur during periods of wakefulness.

Motor adaptation task

In this rotation adaptation task, subjects use a digital tablet (similar to an oversized Palm Pilot) to draw a line displayed on a computer screen from a starting point to a target while overcoming a virtual force 'pushing' the line to one side. As in the previously described tasks, subjects improve over a night with sleep, but not over an equivalent period of daytime wake (Fig. 3g)¹⁶. As with the visual discrimination task's nap studies, training on the task alters subsequent sleep¹⁶, in this case leading to a localized increase in the strength of SWS-associated EEG slow waves recorded over portions of the right parietal cortex implicated in task performance of this task (Fig. 3j)²¹, and this increase correlates with subsequent sleep-dependent improvement in performance ($r = 0.86$, $P < 0.005$; Fig. 3h)⁹. These 0.5–4-Hz EEG oscillations characterize SWS and may themselves facilitate memory consolidation²².

Additional findings and caveats

Studies with other learning tasks help clarify the range and limits of these findings. For example, the serial reaction-time task is a visuo-motor procedural learning task, in which four lights, placed above four response keys, flash in a complex order. As each light flashes, subjects must press the key beneath it. The sequence of lights displayed contains an embedded pattern, which subjects gradually learn. Positron emission tomography (PET) brain imaging studies have demonstrated that on the night after training, regions active during task performance are specifically reactivated during REM²³, a finding similar to the selective regional EEG activation seen during SWS after training on the motor adaptation task¹⁶.

A striking feature of all of the tasks described above is that the enhancement of performance seen over time seems to develop only during sleep. Although one might have expected both wake and sleep to support enhancement, that has not been the case. None of the volunteers showed significant improvement in the task over time awake. But this is not universally true. Training on a complex auditory tone sequence task leads to further improvement after 24 hours even in the absence of sleep²⁴. But, even here, recordings of event-related potentials (ERPs) suggest that subtle sleep-dependent processing is occurring. ERPs measure the EEG activity in response to a specific stimulus, in this case the auditory tone sequence, and allow high temporal resolution of the brain's processing of such stimuli. Here, where no overt sleep-dependent improvement is seen, post-training changes in the ERP, which are thought to be linked to systems-level consolidation processes, are sleep dependent. Thus, some sleep-dependent 'consolidation' seems to be taking place, although it is not reflected in performance gains²⁴. A second auditory discrimination task, in which subjects learn to identify vowel-consonant combinations against a background of white noise, also shows sleep-independent improve-

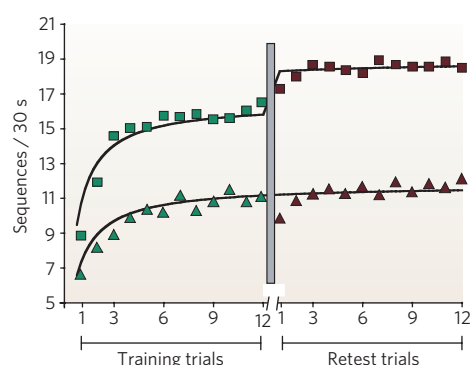


Figure 4 | Fit of data to exponential model of motor learning. Controls ($n = 14$, squares) showed a significant ($P < 0.0001$) 15% overnight improvement, whereas schizophrenia patients ($n = 20$, triangles) showed no improvement ($M < 0.1\%$; $P = 0.998$). The gray bar represents a 24-hour interval between training and retest. Reproduced from ref. 32.

ment, developing over 12 hours, across a day of wake or a night with sleep²⁵.

To make matters more complicated, the serial reaction-time task described above shows very different sleep dependence depending on exactly how the task is implemented²⁶. When subjects are informed at the start of the task that a repeating pattern of stimuli will be presented, they tend to learn the pattern explicitly, within awareness, and can report at least part of the pattern after completion of the protocol. By contrast, when subjects are not so informed, they tend to learn the sequence implicitly, without awareness, and cannot report it. Although both groups show sleep-dependent improvement, the implicit, but not explicit, learners also show improvement across an equivalent period of daytime wake²⁶. Only the explicit version of the task shows exclusively sleep-dependent improvement. Thus it is becoming clear that although we know of no offline procedural performance enhancement that occurs exclusively during periods of wakefulness, some tasks show such improvement during both wake and sleep.

There is one final caveat. Although there is now something close to a consensus that procedural memories are enhanced by sleep (but for a different opinion see ref. 5), there are very few data concerning sleep's role in the stabilization phase of consolidation. Although there is some evidence for stabilization across periods of wake^{20,27}, so far there is no clear evidence of stabilization across periods of sleep.

One important conclusion from this collection of studies is that whether or not sleep-dependent consolidation is observed depends on how training is carried out and how consolidation is measured. Thus interference training on the finger-tapping task blocks sleep-dependent improvements in accuracy, but not in speed²⁰. Post-training sleep deprivation had no effect on post-training enhancement of a complex tone sequence task but does block post-training consolidation measured as ERP changes²⁴. And time-dependent improvement in the serial reaction-time task can be switched from sleep independent to sleep dependent by simply informing subjects that a response pattern exists²⁶.

In summary, procedural learning of perceptual and motor skills are clearly enhanced across periods of sleep, in some cases only across periods of sleep, and most probably only during specific stages of sleep. But what characteristics of a memory control the exclusivity of this sleep-dependent consolidation, as well as which sleep stages are involved, remain unclear, as does the identity of which components of memory consolidation are sleep dependent for any given task.

Clinical considerations

All major psychiatric disorders, the so-called Axis I diagnoses²⁸, such as schizophrenia, bipolar disorder and major depression, have associated sleep disturbances that are considered sequelae to, but never contributors to, the psychiatric condition. But this interaction might be bi-directional. For example, it has been suggested that cognitive deficits

in people with schizophrenia arise in part from an inability to automate simple procedural learning²⁹, a function potentially related to sleep-dependent consolidation. Indeed, chronic, medicated schizophrenia patients show normal practice-dependent improvement in the finger-tapping task discussed above, but show no overnight improvement (Fig. 4)³⁰. Thus, in one major psychiatric illness, at least, one process of sleep-dependent memory consolidation seems to be totally dysfunctional.

Declarative and hippocampal learning

Although sleep-dependent consolidation of procedural learning seems certain, the evidence for declarative memory is weaker. Declarative memory, the formation of which is dependent on the hippocampus and medial temporal lobe, consists primarily of memories for events and facts. Studies of its sleep-dependent consolidation have commonly used tasks such as memorization of nonsense syllables³¹ or word-pair associates³².

As early as 1885, sleep was known to sustain declarative memory more effectively than daytime wake. But this superiority was ascribed to an absence of destructive interference that weakens memories during periods of being awake³¹. Because such memories do not normally show improvement with time, studies measure a balance between stabilization and enhancement on the one hand and interference and passive loss on the other. Until recently, most studies of word-pair associates showed no benefits from sleep (for a review see ref. 32). But a recent series of studies has argued that early night sleep, rich in SWS, supports declarative memory stabilization, and, perhaps, even enhancement³³, although it is not clear what role SWS plays in this consolidation. The stress hormone cortisol is at its lowest early in the night, and infusing cortisol during this time blocks sleep's beneficial effects on these word-pair associates³⁴. Similarly, levels of acetylcholine are lowest early in the night, and consolidation is blocked by cholinesterase inhibitors, which increase acetylcholine levels³⁵. Other studies have found that overnight retention of verbal memory correlates with sleep spindles, an EEG signature of stage 2 NREM^{36,37}. Despite this, others have argued that this sleep benefit merely reflects passive retention in the absence of interfering stimuli during SWS³⁸. Thus, the question of consolidation of word-pair learning remains unresolved.

A second approach to declarative memory has been through the study of hippocampus-dependent spatial memory³⁹. Subjects who train on a virtual navigation task show overnight improvement correlated with increased hippocampal activation during SWS ($r = 0.94$, $P = 0.005$)⁴⁰. This reactivation, measured using PET, parallels a wealth of animal data showing that temporal patterns of activity across networks of hippocampal neurons during performance of a spatial task are repeated during subsequent SWS and REM^{41,42}. In a pattern reminiscent of the human visual discrimination task data¹², replay is seen during SWS hours earlier than in REM. Furthermore, replay during SWS, but not REM, is on a highly compressed timescale. Additional studies have shown REM-dependent consolidation of spatial memory on both radial arm⁴³ and Morris water⁴⁴ mazes in rats. Thus, there is consistent evidence for sleep-dependent consolidation of hippocampus-dependent spatial memory tasks, although relatively few of the data are from humans.

It is striking that, although human studies have implicated SWS in hippocampus-dependent declarative memory consolidation (for examples see refs 33, 40), rat studies suggest that REM is crucial. A possible explanation for this apparent discrepancy comes from a human study suggesting that REM contributes to the consolidation of emotionally charged declarative memories. In this study, recall of emotion, but not neutral, text was enhanced by REM-rich, late-night sleep (emotional, $P < 0.001$; neutral, $P = 0.58$)⁴⁵. Thus, the REM dependence of consolidation in the rat might reflect increased emotional stress.

We thus can draw conclusions concerning declarative and hippocampal memories similar to those for procedural memories, albeit with less confidence. At least some forms of declarative and hippocampus-mediated memory seem to be consolidated across periods of sleep, and, in some cases, preferentially during specific stages of

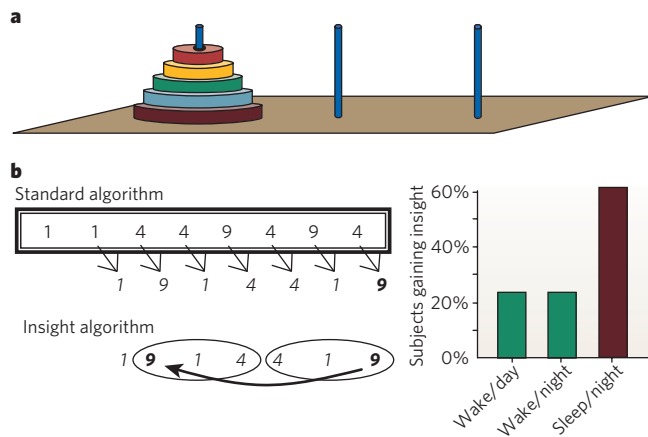


Figure 5 | Complex cognitive procedural learning. **a**, Tower of Hanoi task. Subjects must move one disk at a time, from one pole to another. Disks can be placed only on empty poles or on disks of larger diameter. The goal is to move all the disks to the right-hand pole in the fewest number of moves. For n disks, the optimal score is $2n - 1$ moves. **b**, Mathematical insight. The standard algorithm: subjects are taught a standard algorithm for reducing an eight-digit sequence to a final answer (the bold, italic '9' at the right), through six intermediate calculations (in italics). The 'insight' algorithm: the design of the task is such that the last three calculations form a mirror image of the preceding three, so that the second intermediate calculation always matches the final answer. Right, subjects allowed to sleep between the training and retest showed significantly higher rates of insight compared with those not allowed sleep. Reproduced from refs 49, 50.

sleep. But, again, the characteristics that determine this sleep-dependent consolidation of such memories and the sleep stages involved remain unclear. Even more basic is our uncertainty concerning the effect that sleep has on providing active consolidation, as opposed to merely preventing the negative effects of interference.

Complex cognitive procedural learning

The data presented thus far suggest that the more complex and sophisticated a task, the less confident we are of its sleep-dependent consolidation. Does sleep simply not contribute to more complex forms of human learning? The answer is no. Aside from the fact that motor skills form the basis of dance and music, distinctly human forms of expression, and that more complex motor learning shows more sleep-dependent improvement⁴⁶, several studies suggest that sleep contributes to the consolidation of uniquely human, complex cognitive procedural learning.

The Tower of Hanoi task (Fig. 5a) involves moving five disks from the leftmost pole to the right-most pole, following two rules: first, only the top disk on any pole can be moved at one time, and second, a disk can be moved to a pole only if that pole has no disks or if the disk is smaller than all the disks on the destination pole. When subjects are retested a week after training, a significant ($P = 0.03$) 40% improvement in performance is seen. But if REM is experimentally reduced the night after training, no such improvement is seen.

In another example of complex procedural learning (Fig. 5b)⁴⁷, subjects were taught a complex algorithm for solving a group of mathematical problems. Unknown to the subjects, a simpler solution also exists, which none discovered during training. But at retest, 12 hours later, a subset of subjects discovered this simpler method of performing the task. The probability of discovering it was more than doubled after a night of sleep (Fig. 5b, right, $P = 0.01$)⁴⁷. Thus, sleep enhanced the probability of gaining novel insight into the task, even though subjects did not know that there was an insight to be gained.

These studies suggest a more sophisticated role for sleep in memory consolidation, one that involves the discovery and clarification of complex rules within one's environment, forming and strengthening

associations within memory networks. Such processes are arguably among the most sophisticated human cognitive functions.

Converging evidence

So far, we have focused on human behavioural evidence of sleep-dependent memory consolidation. But several additional lines of investigation provide converging evidence of an important role for sleep in experience-dependent brain plasticity.

Molecular studies

In rats, gene expression during wakefulness and sleep displays state-dependent regulation. High-density microarray studies demonstrate that ~10% of genes expressed in both the cerebral cortex and cerebellum of rats undergo day–night fluctuations, and half of these are specifically dependent on wake–sleep states⁴⁸. Several genes upregulated during sleep, including those encoding calmodulin-dependent protein kinase IV and calcineurin, are believed to contribute to brain plasticity and memory consolidation. Others contribute to the synthesis and maintenance of cell membranes and myelin⁴⁸. In addition, genes are upregulated during sleep in a behaviour-dependent manner. The immediate early gene *zif-268*, which participates in activity-dependent synaptic plasticity, is induced during REM after exposure to enriched environments⁴⁹ (Fig. 6), and experimental induction of hippocampal long term potentiation induces *zif* expression during subsequent REM, first in the amygdala, entorhinal and auditory cortices, and, during subsequent REM periods, in somatosensory and motor cortices⁵⁰.

Cellular studies

Sleep contributes to normal experience-dependent plasticity in both the cat and rat visual systems. In young kittens, monocular deprivation of visual input for 6 hours reduces the responsiveness of cortical neurons to input from the deprived eye. When such deprivation is followed by 6 hours of sleep (but not 6 hours of wake in the dark), the effect is enhanced, producing as much additional change as another 6 hours of active monocular deprivation would⁵¹. Similarly, a crucial period of experience-dependent plasticity in the rat visual system can be extended by REM deprivation⁵².

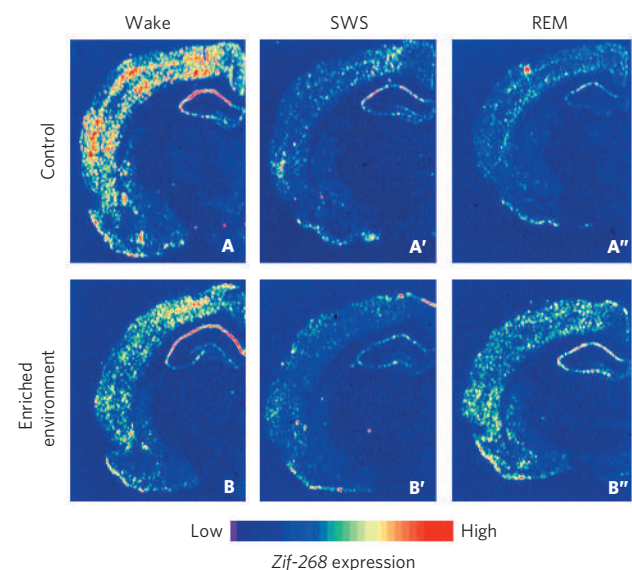


Figure 6 | Experience-dependent upregulation of the synaptic plasticity related immediate early gene *zif-268* during periods of wakefulness, SWS and REM sleep in the rat. Autoradiograms of frontal coronal brain sections in which gene-expression levels best represent the means for each group studied. In controls, *zif-268* expression decreased from wake (A) to SWS (A') and REM (A''). In enriched-environment animals, *zif-268* levels decreased from wake (B) to SWS (B'), but increased from the latter to REM (B''). This effect was particularly noticeable in the cerebral cortex and the hippocampus. Reproduced with permission from ref. 52.

Table 1 | Correlation of sleep-dependent learning and brain measures

Brain measure	Learning/memory task	<i>r</i>	<i>r</i> ²	<i>P</i>	Refs
Quantitative EEG	Motor adaptation	0.86	0.74	<0.005	9
PET	Virtual navigation	0.94	0.88	0.005	62
Time in SWS and REM	Visual discrimination learning	0.89	0.79	<0.0001	5
Time in stage 2	Motor sequence learning	0.72	0.52	<0.01	63
PGO wave density	Shuttle box avoidance (rats)	0.95	0.90	<0.001	64
Spindle density	Paired associates learning	0.56	0.31	<0.05	65
Means and combined probability		0.82	0.69	10 ⁻¹⁵	

Correlation coefficients are between post-sleep improvement and brain measures, including quantitative EEG and REM-related pontogeniculo-occipital (PGO) brain waves, taken during sleep. The combined probability of 10⁻¹⁵ reflects the likelihood of all six studies providing such low probabilities for the null hypothesis.

Neurophysiological studies

As noted earlier, patterns of hippocampal neuronal activation seen during spatial navigation are replayed during subsequent sleep. A similar process of replay is seen in the song system of the Zebra finch, where patterns of neuronal excitation seen during song rehearsal are replayed during sleep⁵³. In this case, sleep leads to a temporary deterioration in song quality⁵⁴, but this deterioration correlates with overall improved learning. This suggests that sleep plays an important role in balancing stabilization of the rehearsed song against the continued plasticity required for further improvement⁵⁴.

Brain imaging studies

Both the visual texture discrimination task^{55,56} and the finger-tapping task⁵⁷ show changes over time in the patterns of regional brain activation associated with task performance, and a night of sleep leads to a reorganization of the brain regions involved in both visual discrimination⁵⁶ and motor sequence⁵⁷ learning. Sleep leads to increased activation in primary visual and motor regions, as expected, but regions of the limbic, frontal and parietal areas also see changes. This suggests that sleep supports enhanced performance by altering the entire strategy used by the brain and allows more automatic execution of the tasks^{24,29}.

Dream studies

An unexpected source of evidence for sleep-dependent memory consolidation comes from the study of dreaming. Dream reports, in common with microarray analyses of gene expression, single-cell and neural network recordings and human brain images, can provide a window into the activity of the sleeping brain, albeit a subjective one. Two such studies^{58,59} suggest, perhaps counterintuitively, that episodic memories are not reactivated during dreaming.

When subjects identified waking sources for dream elements⁵⁸, less than 2% of the elements seemed to be possible replays of episodic memories. In a second study of sleep-onset dreams⁵⁹, subjects who have been playing the computer game Tetris report images from the game, but in a stereotyped form not consistent with a replay of episodic memories. This conclusion is reinforced by similar reports from densely amnesic patients who have no conscious memories of playing the game. Both studies suggest that dream construction occurs without activation of hippocampus-mediated episodic memories; instead they show abstracted images of key elements of the waking events.

Such an absence of episodic memory replay is supported by human PET studies showing that the dorsolateral prefrontal cortex, normally involved in memory recall, is deactivated during sleep, and especially during REM sleep⁶⁰. It is also supported by animal studies suggesting that the flow of information from the hippocampus to the cortex is blocked during REM⁶¹.

Summary

The past 10 years have shown an explosive growth in our knowledge of the relationship between sleep and memory, providing consistent

and strong support for the existence of sleep-dependent memory consolidation. For example, Table 1 summarizes several studies correlating delayed improvement in procedural memory tasks with various measures of intervening sleep. What is noteworthy is that for these six studies, correlation values of 0.56–0.95 were reported, explaining an average across studies of 69% of the variance, a remarkable achievement for any factor contributing to memory consolidation!

Future directions

But several major questions remain. What types of memory are consolidated during sleep? There is strong evidence for some, but not all, forms of procedural learning being consolidated, and more modest evidence for forms of declarative memory, but the rules determining which are consolidated during sleep remain unclear. In addition, what types of consolidation occur during sleep are largely unknown. When sleep-dependent consolidation occurs, both in terms of sleep stages and the time of night, also remains poorly understood. And finally, why these processes are preferentially or exclusively restricted to sleep is completely unknown. The search for answers to these questions is the task of the next decade of research into sleep-dependent memory consolidation. And we still do not know what it means to 'sleep on a problem'. ■

1. Institutes of Oratory. Quintilian <<http://lee.engliastate.edu/11/chapter2.html>> (2004).
2. Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M. & Sagi, D. Dependence on REM Sleep of overnight improvement of a perceptual skill. *Science* **265**, 679–682 (1994).
3. Stickgold, R. & Walker, M. P. Sleep and memory: the ongoing debate. *Sleep* **28**, 1225–1227 (2005).
4. Walker, M. P. & Stickgold, R. Sleep-dependent learning and memory consolidation. *Neuron* **44**, 121–133 (2004).
5. Vertes, R. P. & Siegel, J. M. Time for the sleep community to take a critical look at the purported role of sleep in memory processing. *Sleep* **28**, 1228–1229 (2005).
6. Schacter, D. L. & Tulving, E. *Memory Systems* (MIT Press, Cambridge, MA, 1994).
7. Müller, G. E. & Pilzecker, A. Experimentelle Beiträge zur Lehre vom Gedächtnis. *Z. Psychol.* **1**, 1–288 (1900).
8. Duncan, C. P. The retroactive effect of electroshock on learning. *J. Comp. Physiol. Psychol.* **42**, 32–34 (1949).
9. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
10. Dudai, Y. The neurobiology of consolidations, or, how stable is the engram? *Annu. Rev. Psychol.* **55**, 51–86 (2004).
11. Stickgold, R. Sleep: off-line memory reprocessing. *Trends Cogn. Sci.* **2**, 484–492 (1998).
12. Stickgold, R., Whidbee, D., Schirmer, B., Patel, V. & Hobson, J. A. Visual discrimination task improvement: A multi-step process occurring during sleep. *J. Cogn. Neurosci.* **12**, 246–254 (2000).
13. Stickgold, R., James, L. & Hobson, J. A. Visual discrimination learning requires post-training sleep. *Nature Neurosci.* **2**, 1237–1238 (2000).
14. Walker, M., Brakefield, T., Morgan, A., Hobson, J. A. & Stickgold, R. Practice with sleep makes perfect: Sleep dependent motor skill learning. *Neuron* **35**, 205–211 (2002).
15. Fischer, S., Hallschmid, M., Elsner, A. L. & Born, J. Sleep forms memory for finger skills. *Proc. Natl Acad. Sci. USA* **99**, 11987–11991 (2002).
16. Huber, R., Ghilardi, M. F., Massimini, M. & Tononi, G. Local sleep and learning. *Nature* **430**, 78–81 (2004).
17. Gais, S., Pilhal, W., Wagner, U. & Born, J. Early sleep triggers memory for early visual discrimination skills. *Nature Neurosci.* **3**, 1335–1339 (2000).
18. Walker, M. P. & Stickgold, R. It's practice, with sleep, that makes perfect: Implications of sleep-dependent learning and plasticity for skill performance. *Clinics Sports Med.* (in the press).
19. Walker, M. P. *et al.* Sleep and the time course of motor skill learning. *Learn. Mem.* **10**, 275–284 (2003).
20. Walker, M. P., Brakefield, T., Hobson, J. A. & Stickgold, R. Dissociable stages of human memory consolidation and reconsolidation. *Nature* **425**, 616–620 (2003).
21. Ghilardi, M. *et al.* Patterns of regional brain activation associated with different forms of motor learning. *Brain Res.* **871**, 127–145 (2000).
22. Steriade, M. Active neocortical processes during quiescent sleep. *Arch. Ital. Biol.* **139**, 37–51 (2001).
23. Maquet, P. *et al.* Experience-dependent changes in cerebral activation during human REM sleep. *Nature Neurosci.* **3**, 831–836 (2000).
24. Atienza, M., Cantero, J. L. & Stickgold, R. Posttraining sleep enhances automaticity in perceptual discrimination. *J. Cogn. Neurosci.* **16**, 53–64 (2004).
25. Roth, D. A., Kishon-Rabin, L., Hildesheimer, M. & Karni, A. A latent consolidation phase in auditory identification learning: Time in the awake state is sufficient. *Learn. Mem.* **12**, 159–164 (2005).
26. Robertson, E. M., Pascual-Leone, A. & Press, D. Z. Awareness modifies the skill-learning benefits of sleep. *Curr. Biol.* **14**, 208–212 (2004).
27. Brashers-Krug, T., Shadmehr, R. & Bizzi, E. Consolidation in human motor memory. *Nature* **382**, 252–255 (1996).
28. American Psychological Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn (American Psychiatric Association, Washington DC, 1994).
29. Manoach, D. S. Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. *Schizophr. Res.* **60**, 285–298 (2003).

30. Manoach, D. S. *et al.* A failure of sleep-dependent procedural learning in chronic, medicated schizophrenia. *Biol. Psych.* **56**, 951–956 (2004).
31. Ebbinghaus, H. *Über das Gedächtnis* (Teachers' College, New York, 1885).
32. Smith, C. Sleep states and memory processes in humans: procedural versus declarative memory systems. *Sleep Med. Rev.* **5**, 491–506 (2001).
33. Plihal, W. & Born, J. Effects of early and late nocturnal sleep on declarative and procedural memory. *J. Cogn. Neurosci.* **9**, 534–547 (1997).
34. Plihal, W. & Born, J. Memory consolidation in human sleep depends on inhibition of glucocorticoid release. *Neuroreport* **10**, 2741–2747 (1999).
35. Gais, S. & Born, J. Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proc. Natl Acad. Sci. USA* **101**, 2140–2144 (2004).
36. Schabus, M. *et al.* Sleep spindles and their significance for declarative memory consolidation. *Sleep* **27**, 1479–1485 (2004).
37. Clemens, Z., Fabo, D. & Halasz, P. Overnight verbal memory retention correlates with the number of sleep spindles. *Neuroscience* **132**, 529–535 (2005).
38. Wixted, J. T. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* **55**, 235–269 (2004).
39. Plihal, W. & Born, J. Effects of early and late nocturnal sleep on priming and spatial memory. *Psychophysiology* **36**, 571–582 (1999).
40. Peigneux, P. *et al.* Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* **44**, 535–545 (2004).
41. Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145–156 (2001).
42. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
43. Smith, C. T., Conway, J. M. & Rose, G. M. Brief paradoxical sleep deprivation impairs reference, but not working, memory in the radial arm maze task. *Neurobiol. Learn. Mem.* **69**, 211–217 (1998).
44. Smith, C. & Rose, G. M. Evidence for a paradoxical sleep window for place learning in the Morris water maze. *Physiol. Behav.* **59**, 93–97 (1996).
45. Wagner, U., Gais, S. & Born, J. Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learn. Mem.* **8**, 112–119 (2001).
46. Kuriyama, K., Stickgold, R. & Walker, M. P. Sleep-dependent learning and motor skill complexity. *Learn. Mem.* **11**, 705–713 (2004).
47. Wagner, U., Gais, S., Haider, H., Verleger, R. & Born, J. Sleep inspires insight. *Nature* **427**, 352–355 (2004).
48. Cirelli, C., Gutierrez, C. M. & Tononi, G. Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron* **41**, 35–43 (2004).
49. Ribeiro, S., Goyal, V., Mello, C. V. & Pavlides, C. Brain gene expression during REM sleep depends on prior waking experience. *Learn. Mem.* **6**, 500–508 (1999).
50. Ribeiro, S. *et al.* Induction of hippocampal long-term potentiation during waking leads to increased extrahippocampal zif-268 expression during ensuing rapid-eye-movement sleep. *J. Neurosci.* **22**, 10914–10923 (2002).
51. Frank, M. G., Issa, N. P. & Stryker, M. P. Sleep enhances plasticity in the developing visual cortex. *Neuron* **30**, 275–287 (2001).
52. Shaffery, J. P., Sinton, C. M., Bissette, G., Roffwarg, H. P. & Marks, G. A. Rapid eye movement sleep deprivation modifies expression of long-term potentiation in visual cortex of immature rats. *Neuroscience* **110**, 431–443 (2002).
53. Dave, A. S. & Margoliash, D. Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* **290**, 812–816 (2000).
54. Derégnaucourt, S., Mitra, P. P., Feher, O., Pytte, C. & Tchernichovski, O. How sleep affects the developmental learning of bird song. *Nature* **433**, 710–716 (2005).
55. Schwartz, S., Maquet, P. & Frith, C. Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. *Proc. Natl Acad. Sci. USA* **99**, 17137–17142 (2002).
56. Walker, M. P., Stickgold, R., Jolesz, F. A. & Yoo, S.-S. The functional anatomy of sleep-dependent visual skill learning. *Cerebral Cortex* published online 9 February, 2005 (doi:10.1093/cercor/bhi043).
57. Walker, M. P., Stickgold, R., Alsop, D., Gaab, N. & Schlaug, G. Sleep-dependent motor memory plasticity in the human brain. *Neuroscience* **133**, 911–917 (2005).
58. Fosse, M. J., Fosse, R., Hobson, J. A. & Stickgold, R. J. Dreaming and episodic memory: a functional dissociation? *J. Cogn. Neurosci.* **15**, 1–9 (2003).
59. Stickgold, R., Malia, A., Maguire, D., Roddenberry, D. & O'Connor, M. Replaying the game: Hypnagogic images in normals and amnesiacs. *Science* **290**, 350–353 (2000).
60. Maquet, P. Functional neuroimaging of normal human sleep by positron emission tomography. *J. Sleep Res.* **9**, 207–231 (2000).
61. Buzsáki, G. The hippocampo-neocortical dialogue. *Cerebral Cortex* **6**, 81–92 (1996).
62. Peigneux, P. *et al.* Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* **44**, 535–545 (2004).
63. Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A. & Stickgold, R. Practice with sleep makes perfect: sleep dependent motor skill learning. *Neuron* **35**, 205–211 (2002).
64. Datta, S. Avoidance task training potentiates phasic pontine-wave density in the rat: A mechanism for sleep-dependent plasticity. *J. Neurosci.* **20**, 8607–8613 (2000).
65. Gais, S., Molle, M., Helms, K. & Born, J. Learning-dependent increases in sleep spindle density. *J. Neurosci.* **22**, 6830–6834 (2002).

Acknowledgements This work was supported by grants from the US National Institutes of Health.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to R.S. (rstickgold@hms.harvard.edu).

Insights from studying human sleep disorders

Mark W. Mahowald¹ and Carlos H. Schenck¹

Problems with sleep are one of the commonest reasons for seeking medical attention. Knowledge gained from basic research into sleep in animals has led to marked advances in the understanding of human sleep, with important diagnostic and therapeutic implications. At the same time, research guided by human sleep disorders is leading to important basic sleep concepts. For example, sleep may not be a global, but rather a local, brain phenomenon. Furthermore, contrary to common assumptions, wakefulness, rapid eye movement (REM) and non-REM sleep are not mutually exclusive states. This striking realization explains a fascinating range of clinical phenomena.

Wake/sleep complaints are second only to complaints of pain as a cause to seek medical attention. Undiagnosed and untreated wake/sleep complaints exact an enormous toll at the personal level in terms of misery and at the societal level in socioeconomic consequences. Knowledge of sleep and sleep disorders has markedly increased over the past half-century, spearheaded by the discovery that sleep is far more than the passive absence of wakefulness. Not only is sleep an active brain process, but it actually represents two completely different states of being: non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. These states are as different from each other as each is from wakefulness.

Most parts of the brain are active during all three states of being — wakefulness, NREM sleep and REM sleep — but active in a different mode, reflecting a striking reorganization of brain activity and function during these different states. Knowledge gained from basic research on sleep in animals has led to amazing advances in the understanding of human sleep, with important diagnostic and effective therapeutic implications. In addition, research guided by human sleep disorders is leading to important basic sleep concepts. These include the ideas that sleep may not be a global, but rather a local, brain phenomenon, and that wakefulness, NREM sleep and REM sleep are not mutually exclusive states: state dissociation or admixture might explain fascinating clinical phenomena. In Fig. 1, a period of 'ambiguous' sleep, with the simultaneous appearance of elements of both REM and NREM sleep, can be seen.

Although there are nearly 100 identified sleep/wake disorders, most sleep complaints conveniently fall into four categories: hypersomnia (excessive daytime sleepiness without obvious explanation), insomnia (trouble falling or staying asleep), circadian rhythm (biological clock) disorders, and parasomnias (complex behaviours arising from the sleep period). This review will focus on key disorders in each category that have provided important and exciting insights to our understanding of the wake/sleep process.

Hypersomnia

Hypersomnia should be taken very seriously, as sleepiness from any cause results in impaired sustained attention, with adverse, occasion-

ally disastrous, socioeconomic consequences in the classroom, workplace or on the highways. For instance, it is likely that more than 100,000 motor vehicle crashes annually in the United States are caused by driving while drowsy. Major disasters such as Three Mile Island, *Exxon Valdez*, Bhopal and *Challenger* were all officially attributed to sleepiness-related impaired judgement in the workplace¹. The most common cause of hypersomnia is volitional sleep deprivation for social or economic reasons (we get 20% less sleep than previous generations, and there is no evidence that earlier generations required more sleep, or that ours needs less). Sleepiness not explained by volitional sleep deprivation is almost always due to an underlying sleep disorder, most commonly obstructive sleep apnoea or narcolepsy.

Obstructive sleep apnoea

Obstructive sleep apnoea (OSA) is the most common medical disorder causing hypersomnia, affecting over 2% of adult women and 4% of adult men. It is seen primarily in people who are loud snorers and is characterized by collapse of the upper airway during sleep. This upper airway collapse may be associated with a fall in the blood oxygen level and results in repetitive arousals (up to 100 per hour of sleep) to re-establish upper airway airflow. These brief arousals are not perceived by the individual but result in excessive daytime sleepiness. Important insights into OSA have been that, contrary to initial reports, OSA is not confined to middle-aged, overweight males but may be seen in children (3% of all children), women and thin individuals. The fact that OSA is a risk factor for hypertension and has been associated with heart disease and type 2 diabetes is of particular concern given the high prevalence of OSA and that obesity (currently an epidemic in many countries) is one risk factor for the development of OSA².

Narcolepsy

Narcolepsy is a relatively rare neurological disorder affecting 1 in 2,000 individuals. It is characterized by the tendency to fall asleep inappropriately during the daytime, particularly during sedentary or non-stimulating activities, despite having obtained an adequate amount of sleep the preceding night. Narcolepsy is the only known disorder representing abnormalities of the wake/sleep generators. Other symptoms

¹Minnesota Regional Sleep Disorders Center, and Department of Neurology (MWM) and Psychiatry (CHS) Hennepin County Medical Center, and University of MN Medical School, Minneapolis, Minnesota 55415, USA.

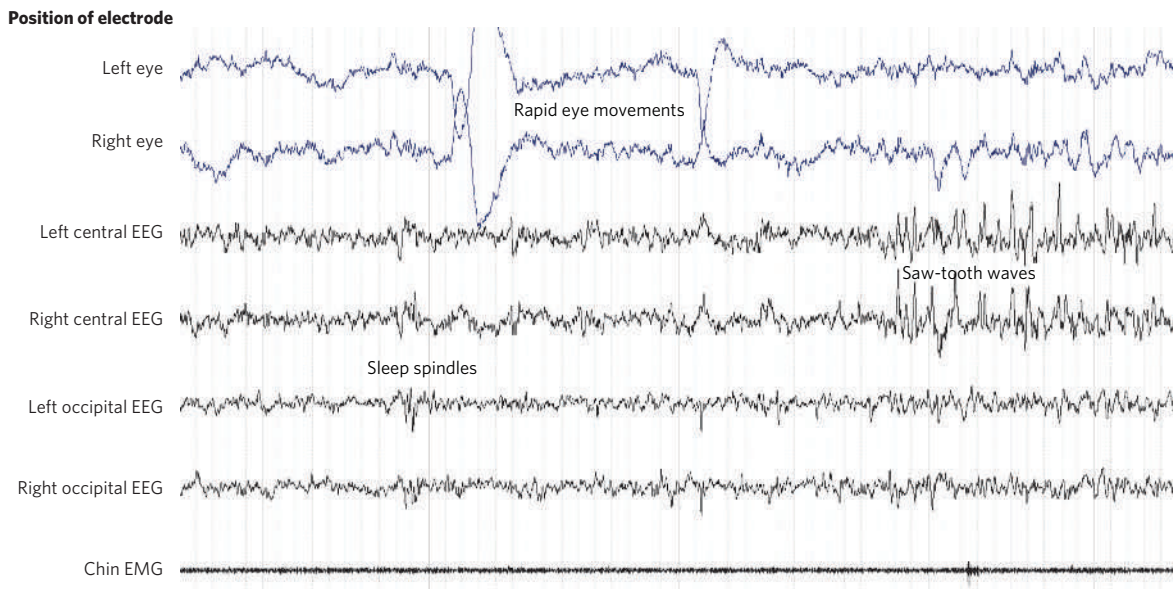


Figure 1 | Ambiguous sleep in a patient with narcolepsy. Note the simultaneous occurrence of elements of NREM sleep (sleep spindles) and REM sleep (rapid eye movements, saw-tooth waves and muscle atonia). EEG, electroencephalogram; EMG, electromyogram.

of narcolepsy include: first, cataplexy (sudden brief spells of muscle weakness), often triggered by emotionally laden events; second, hypnagogic (occurring at sleep onset) or hypnopompic (occurring at sleep offset) hallucinations; third, sleep paralysis (awakening to find the entire body paralyzed, with the exception of being able to breathe and move the eyes); fourth, automatic behaviour; and last, disrupted night-time sleep³.

The study of narcolepsy has been instrumental in promoting the concept of state dissociation: wakefulness and sleep are not necessarily mutually exclusive states, and furthermore, elements of one state of being may intrude inappropriately into another, often with striking consequences. Importantly, people with narcolepsy do not sleep more per 24 hours than non-narcoleptics; they are unable to keep the normal boundaries of wakefulness, NREM sleep and REM sleep⁴. The automatic behaviour (driving past the desired freeway exit, putting clothing into the refrigerator) represents an admixture of wakefulness and NREM sleep. There is enough wakefulness to perform complex behaviour but not enough for conscious awareness of the behaviours.

Sleep paralysis and cataplexy represent the simultaneous occurrence of REM sleep-related muscle paralysis and wakefulness. If the paralysis intrudes into wakefulness, the result is cataplexy. If it persists into wakefulness from a period of REM sleep, sleep paralysis results. The waking hallucinations represent the release of sleep-related dreaming into wakefulness, and the disrupted night-time sleep is a manifestation of the 'state-boundary dyscontrol' aspect of narcolepsy. Sleep paralysis and hypnagogic/hypnopompic hallucinations, but not cataplexy, are often experienced by people who do not have narcolepsy, particularly in the setting of sleep deprivation.

There was a time when narcolepsy was felt to be a psychiatric condition. There were somewhat plausible psychodynamic explanations (primarily as avoidance behaviours) for the symptoms of sleepiness, cataplexy and sleep paralysis. But narcolepsy is clearly caused by abnormalities of the central nervous system.

Narcolepsy has a clear genetic component, with over 90% of individuals with narcolepsy carrying the *HLA-DR2/DQ1* (under current nomenclature *HLA-DRI5* and *HLA-DQ6*) gene (found in less than 30% of the general population). It is currently felt that DQ6, which corresponds at the genomic level to the subregions DQB1*0602 and DQA1*0102 on chromosome 6, is one of the more reliable markers for narcolepsy across ethnic groups⁵. This association is present in the different ethnic populations to varying degrees and represents the highest disease-HLA linkage known in medicine. Despite a genetic

component, the risk of a first-degree relative developing narcolepsy is only 1–2%, but that represents a 10–40-fold increase compared with the general population⁶. Clearly, there is a genetic component. However, that component is neither necessary nor sufficient to cause narcolepsy.

One of the most important and exciting discoveries in the field of sleep medicine resulted from animal experiments into narcolepsy and was the identification of the relationship between hypocretin-1 (also known as orexin) and narcolepsy. Hypocretin-1 is a neuropeptide confined to a small number of cells in the hypothalamus. It seems that patients with narcolepsy have lost these hypocretin-producing cells, possibly through an immune-mediated mechanism⁷. Undetectable levels of hypocretin-1 in the cerebrospinal fluid (CSF) are very specific for patients with narcolepsy who have cataplexy and who are HLA DQB1*0602 positive. Absent CSF hypocretin-1 levels have not been found in any other conditions that could be confused clinically with narcolepsy, and this suggests that CSF hypocretin determinations may be of value in the diagnosis of narcolepsy in difficult cases⁸.

There are numerous effective treatments for narcolepsy such as stimulant medications. Discovery of the role of hypocretin deficiency in narcolepsy will undoubtedly lead to more specific targeted treatments for narcolepsy and will teach much about wake/sleep function in general. It has already been shown that systemic administration of hypocretin-1 in narcoleptic dogs reduced cataplexy and normalized sleep and waking durations⁹.

Insomnia

Insomnia is the most prevalent sleep complaint in the general population. It is not defined by total sleep time, but, rather, by the inability to obtain sleep that is sufficiently long or 'good enough' to result in feeling rested or restored the following day. Although insomnia may be due to many underlying medical, psychiatric or psychological conditions, there is growing evidence that some insomnia is constitutional in nature. Many people with insomnia do not have any identifiable psychiatric or psychological problems. Furthermore, there is evidence that untreated insomnia is a risk factor for the development of psychiatric problems such as depression or substance abuse. Importantly, the relationship between insomnia and psychiatric conditions is bi-directional: depression may cause insomnia, and insomnia may cause depression.

There is convincing evidence that insomniacs may be in a constant state of hyperarousal. Many are actually less sleepy during the day than

non-insomniacs as measured by objective daytime nap studies. And they also have an increase in metabolic rate across the 24-hour period¹⁰. It has been proposed that chronic insomniacs may suffer from a more general disorder of hyperarousal that may be responsible for both the daytime symptoms and poor nocturnal sleep¹¹. Insomniacs experience an overall increase of adrenocorticotrophic hormone (ACTH) and cortisol secretion¹². The increased arousal level shown in electroencephalogram (EEG) power-spectra studies could help explain the impairment of the perception of having slept experienced by many insomniacs¹³. Neuroimaging studies in insomnia lend further support to underlying physiological abnormalities¹⁴.

Behavioural treatment for insomnia can be quite effective but can be time-consuming. Numerous pharmacological treatments are available. Over-the-counter sleep aids are of very questionable benefit¹⁵. Commonly, medications used to treat depression are prescribed to treat insomnia, although there is very little evidence that antidepressant medications are effective in the treatment of insomnia that is not associated with depression¹⁶. There are two classes of medication approved for the treatment of insomnia: the benzodiazepines and the newer, non-benzodiazepines. The risks of tolerance, abuse and dependency associated with chronic benzodiazepine administration for patients with well-documented sleep disorders have been greatly overrated¹⁷. Combined behavioural and pharmacological treatments are often effective¹⁸. Benzodiazepines may be safely and effectively administered for longer than three weeks¹⁹.

There has been much interest in using melatonin as a sleep-inducing agent. Melatonin is secreted by the pineal gland, and its secretion is controlled by the light–dark cycle. Although often thought of as a ‘sleep’ hormone, it is actually a ‘darkness’ hormone, being secreted at night in both day-active and night-active species. Its effect on sleep is variable. Melatonin may prove useful in certain specific conditions. For example, melatonin is effective in improving sleep in a subgroup of elderly patients with insomnia who have low melatonin levels. Melatonin is clearly ineffective in a sizable number of those suffering from insomnia, and its use in the general treatment of insomnia remains to be clarified²⁰. Melatonin agonists are being evaluated as sleep-inducing agents.

Awareness of the neurophysiological factors that predispose to insomnia and the mechanism of action of sedative-hypnotic agents are important elements in the diagnosis and successful management of insomnia. Further study of insomnia in neurological conditions will expand our understanding of sleep/wake patterns and brain function. If some patients with insomnia are physiologically hyperaroused, a case may be made for chronic administration of sedative-hypnotic agents. The recent identification of a number of sleep/wake-related neurotransmitters and neuropeptides will lead to the development of newer and more selective sleep-promoting agents.

Restless legs syndrome

Restless legs syndrome (RLS) is one of the most common causes of severe insomnia. It is a neurological sensory/movement disorder affecting 5–15% of the general population²¹. RLS is characterized primarily by a vague and difficult-to-describe unpleasant sensation in the legs. This discomfort appears primarily during periods of inactivity, particularly during the transition from wake to sleep in the evening. Patients often have difficulty in describing the unpleasant sensations; they rarely use conventional terms of discomfort such as ‘numbness, tingling or pain’, but rather bizarre terms such as ‘pulling, searing, drawing, crawling, shimmering or boring’, suggesting that RLS sensations are unlike any experienced by unaffected individuals. These distressing sensations are typically relieved only by movement or stimulation of the legs. Many different techniques have been found by patients: walking about, stomping the feet, rubbing, squeezing or stroking the legs, taking hot showers or baths, or applying ointment, hot packs or wraps to the legs. Although these treatments are effective while they are being performed, the discomfort usually returns as soon as the individual becomes inactive or returns to bed to try to sleep. The

Box 1 | How much sleep do you need?

- The total sleep requirement is genetically determined on an individual basis.
- Although the average is 7.5–8 hours per night, the range is between 4 hours and 10–11 hours.
- One is sleep-deprived if:
 - One uses an alarm clock to be awakened in the morning (otherwise, the brain would have awakened before the alarm went off when it had accumulated as much sleep as it needs).
 - One tends to fall asleep during periods of reduced environmental stimulation. Boring lectures, dimly lit rooms, heavy meals or long automobiles drives do not cause sleepiness, they simply unmask it.
- The best way to determine one's constitutional sleep requirement is to recall (or speculate as to) what one's would be if one were in an environment where the wake-sleep pattern was completely free of external demands or constraints from work, family, school, social or meals.

motor restlessness often appears to follow a circadian pattern, peaking between midnight and 04:00 (ref. 22). The higher prevalence of depression and anxiety found associated with RLS is felt to be secondary to the RLS²³.

Recent studies suggest there may be a susceptibility gene locus, which would explain why RLS is often familial²⁴. RLS is commonly seen in pregnancy, haemodialysis or peritoneal dialysis for renal failure and iron-deficiency anaemia. Its relationship with iron metabolism abnormalities has led to studies indicating that RLS is associated with abnormal iron metabolism within the central nervous system²⁵. Recent elegant functional neuroimaging studies have identified thalamic, red nucleus and brainstem involvement in the generation of periodic limb movements in patients with RLS²⁶. One PET study found reduced dopamine 2 binding in the caudate and putamen²⁷. A subcortical origin of RLS is supported by transcranial magnetic stimulation studies²⁸.

Demonstration of a continuous hypersensitivity to pin-prick, but not to light touch, confirms a sensory component to RLS²⁹ and may explain the efficacy of opiate medications in RLS³⁰.

Most cases of RLS respond gratifyingly to medical treatment including anti-parkinsonian agents, benzodiazepines, opiates and anti-convulsant medication. The dopaminergic anti-parkinsonian medications are particularly effective²⁴. Further study of the relationship between dopamine and central nervous system iron metabolism will be of great interest and value to neurophysiologists, neurochemists, neuropharmacologists and, of course, patients.

Circadian rhythm disorders

Most living creatures follow a relentless and pervasive daily rhythm of activity and rest that is ultimately linked to the geophysical light–dark cycle. Plants, animals, even unicellular organisms, show daily variations in metabolic activity, locomotion, feeding and many other functions³¹. The importance of the light–dark cycle upon the human biological clock is underscored by the fact that in totally blind humans, only one-third will be entrained to the environment. One-third will have a cycle that is 24 hours but out of phase with the environment, with the remaining third experiencing a free-running pattern longer than 24 hours³².

The primary symptom of circadian rhythm disorders is the inability to sleep during the desired sleep time. Once asleep, there is no abnormality of sleep, only an abnormality of the timing of sleep. The cause of all circadian rhythm disorders is an inability of the individual's biological clock to adjust to the demands of the geophysical environment. Wake–sleep schedule disorders fall into two categories: primary (malfunction of the biologic clock *per se*) and secondary (resulting from environmental effects upon the underlying clock). The secondary disorders (such as jet-lag and shift work) are usually immediately apparent upon simple questioning of the patient. The primary disorders may be much more difficult to diagnose, as they typically masquerade as other sleep, medical or psychiatric disorders such as hypersomnia, insomnia, substance (sedative-hypnotic or stimulant)



Figure 2 | A sleepwalking episode. A 23-year-old female demonstrates a very rapid progression from lying in her bed in deep sleep (slow-wave NREM sleep) to have a precipitous arousal with immediate standing up and sleepwalking that culminates with her tripping and nearly falling out of bed.

abuse or psychiatric conditions.

Until recently, the circadian rhythm disorders were only of academic interest, as no effective treatments existed. Fortunately, this has changed, and many patients with these often incapacitating disorders will benefit from accurate diagnosis and appropriate treatment. The mainstays of treatment are chronotherapy and phototherapy. In addition, there are promising new pharmacological treatments on the horizon³³.

In chronotherapy, the desirable total sleep time is determined by sleep logs during a 'free-running' period. The patient then delays or advances sleep onset, by a few hours every day and sleeping only the predetermined number of hours until the sleep onset time is at the desired time, at which time the patient attempts to maintain that time³⁴. This method requires several days of free time and can be difficult if sleeping quarters cannot be kept dark and quiet when sleep will be occurring during the daytime.

Phototherapy (exposure to bright light) has a potent effect upon the biological clock, and exposure at strategic times of the wake/sleep cycle results in a change in the underlying rhythm. This has afforded an opportunity to treat circadian dysrhythmias effectively. The timing and duration of the phototherapy depend upon diagnosis and individual response. The patient sits at a prescribed distance from a bright light producing 2,500 lux at the point where it reaches the patient. The effect of light upon human rhythms varies with intensity, wavelength, timing and duration of exposure. Once the desired sleep period time has been achieved, continued light exposure must be maintained. Much remains to be learned regarding these variables and the effectiveness of phototherapy in the clinical setting³⁵.

Delayed sleep syndrome

In delayed sleep-phase syndrome (DSPS), the patient falls asleep late and rises late. There is a striking inability to fall asleep at an earlier, more desirable time. This may present itself as either sleep-onset insomnia or daytime hypersomnia (particularly in the morning). DSPS is the most common of the primary circadian dysrhythmias, and may, in part, be the consequence of societal increases in night-time activities³⁶. For example, a college student is habitually unable to fall asleep until 02:00, and has great difficulty getting up in time for her 08:00 classes from Monday to Friday. She finds herself dozing off during morning classes. On Saturday and Sunday she sleeps in until about 10:00, and feels rested upon arising, with no episodes of dozing during the day. Combinations of chronotherapy, phototherapy and medications may be effective in 'resetting' the clock. Unfortunately, the treatment regimen must be maintained, or the clock will again become delayed.

Advanced sleep-phase syndrome

Individuals suffering from advanced sleep-phase syndrome (ASPS) fall asleep early and awaken earlier than desired. They are unable to remain awake until the desired time, falling asleep in the early evening and awakening in the very early hours of the morning. This may present as hypersomnia (particularly in the evening) or sleep-mainte-

nance insomnia. Patients complain of interruption of evening activities by their sleepiness. They may avoid evening social activities, fearing the intrusive sleepiness. The undesirable early morning awakenings in this condition may lead to a misdiagnosis of depression. Bright light exposure in the evening may delay the clock to a more acceptable pattern.

Other, less common, circadian dysrhythmias beyond the scope of this review include a 'non-24-hour wake-sleep pattern' and 'irregular wake-sleep pattern'.

Underscoring the inherent nature of these conditions, human genetic studies have identified specific genes associated with both advanced and delayed sleep-phase syndromes^{37–39}.

Drugs that shift biological rhythms are called chronobiotics. One promising pharmacological treatment is melatonin. Melatonin is secreted by the pineal gland, and its secretion is coupled to the wake-sleep cycle and to the circadian cortisol rhythm. It is a valuable marker of the underlying wake-sleep period. It is likely that melatonin plays an important role in biological rhythms. There is evidence that administration of exogenous melatonin may alter the biological rhythm in certain conditions^{33,40}.

Circadian dysrhythmias are common, and may be disabling in terms of academic, employment or family consequences, and so the sleep medicine field is eagerly awaiting the development of effective chronobiotics.

Parasomnias

Parasomnias are defined as unpleasant or undesirable behavioural or experiential phenomena that occur predominantly or exclusively during sleep. They were initially thought to be a single phenomenon, often attributed to psychiatric disease. It is now clear that parasomnias are not a unitary phenomenon but rather are the manifestation of a wide variety of completely different conditions, most of which are diagnosable and treatable. The common parasomnias are another example of 'dissociated sleep states', representing the simultaneous admixture of wakefulness and either NREM sleep (disorders of arousal such as sleepwalking or sleep terrors) or wakefulness and REM sleep (REM sleep behaviour disorder). Parasomnias may result in striking clinical phenomena, which appear as the brain becomes reorganized across states, and are therefore apt to occur during the transition from one state to another.

In view of the large number of neural networks, neurotransmitters and other state-determining substances that must be recruited synchronously for full state declaration and the frequent transitions among states during the wake-sleep cycle, it is surprising that errors in state declaration do not occur more frequently than they do⁴¹. In addition to the phenomenon of state dissociation, in which two states of being overlap or occur simultaneously, there are probably additional underlying physiological phenomena that contribute to the appearance of complex motor behaviours during sleep. These include: activation of locomotor centres during sleep, sleep inertia (a period of confusion or disorientation during the transition from sleep to wakefulness) upon arousal and sleep state instability (oscillation between



Figure 3 | The rapid evolution of a sleep terror episode. A 24-year-old male patient has a long and increasingly severe history of violent sleep terrors. In the video prints shown here, the patient emerged precipitously from slow-wave NREM sleep, without any physiological or motor forewarning. This is typical of sleep terrors (see the text for more details).

wakefulness and sleep)⁴².

The parasomnias, which are usually due to admixtures of states of being, along with narcolepsy, support the concept that wake and sleep are not mutually exclusive states, and that sleep is not necessarily a global brain phenomenon.

NREM parasomnias

The disorders of arousal are the most impressive and most common of the NREM sleep parasomnias. These share common features. They tend to arise from slow-wave sleep (stages 3 and 4 of NREM sleep) and therefore usually occur during the first third of the sleep cycle (and rarely during naps). They are common in childhood, usually decreasing in frequency with increasing age^{43,44}.

Disorders of arousal may be triggered by febrile illness, alcohol, previous sleep deprivation, physical activity, emotional stress or medications. Such precipitants should be thought of as triggering events in susceptible individuals and not the cause. Persistence of these behaviours beyond childhood or their development in adulthood has erroneously been taken as an indication of underlying psychopathology. Careful clinical evaluations of patients with disorders of arousal have dispelled the myth that these conditions are the manifestation of significant underlying psychiatric or psychological problems^{45–47}.

Disorders of arousal occur on a broad spectrum ranging from confusional arousals, through somnambulism (sleep walking) to sleep terrors. Some take the form of 'specialized' behaviours such as sleep-related eating and sleep-related sexual activity^{48–51}.

Confusional arousals

These are often seen in children and are characterized by movements in bed, occasionally thrashing about or inconsolable crying⁵². 'Sleep drunkenness' is probably a variation on this theme⁵³. The prevalence of confusional arousals in adults is approximately 4% (ref. 54).

Sleepwalking

Sleepwalking is prevalent in childhood (1–17%), peaking at 11–12

years of age, and is far more common in adults (nearly 4%) than generally acknowledged^{54,55}. Sleepwalking may be either calm or agitated, with varying degrees of complexity and duration (Fig. 2).

Sleep terrors

The sleep terror is the most dramatic disorder of arousal, frequently initiated by a loud, blood-curdling scream associated with extreme panic, followed by prominent motor activity such as hitting the wall, running around or out of the bedroom, even out of the house, resulting in bodily injury or property damage. A universal feature is inconsolability, and attempts at consolation are fruitless and may serve only to prolong or even intensify the confusional state. Complete amnesia for the activity is typical, but may be incomplete⁵⁶. The intense endogenous arousal and exogenous unarousability constitute a curious paradox. As with sleepwalking, sleep terrors are much more prevalent in adults than generally acknowledged (up to 3%) (ref. 54). Although usually benign, these behaviours may be violent (Fig. 3), resulting in considerable injury to the victim or others or damage to the environment, occasionally with forensic implications⁵⁷. Treatment options include reassurance, behavioural or pharmacological approaches⁵⁸.

REM sleep parasomnias

The most common and best-studied REM sleep parasomnia is the REM sleep behaviour disorder (RBD). In patients with RBD, somatic muscle atonia, one of the defining features of REM sleep, is absent, permitting the acting out of dream mentation, often with violent or injurious results (Fig. 4). Figure 5 demonstrates the striking release of motor activity during REM sleep in a patient with RBD.

The cases so far reported indicate strikingly similar clinical features⁵⁹. The presenting complaint is that of vigorous sleep behaviours usually accompanying vivid dreams. These behaviours may result in repeated injury, including ecchymoses, lacerations and fractures. Some of the self-protection measures taken by the patients (tethering themselves to the bed, using sleeping bags or pillow barricades, or sleeping on a mattress in an empty room) reveal the recurrent and serious



Figure 4 | REM sleep behaviour disorder. Pictures taken from a video in a sleep laboratory of older men who have RBD. These men throw punches while dreaming of fighting during REM sleep.

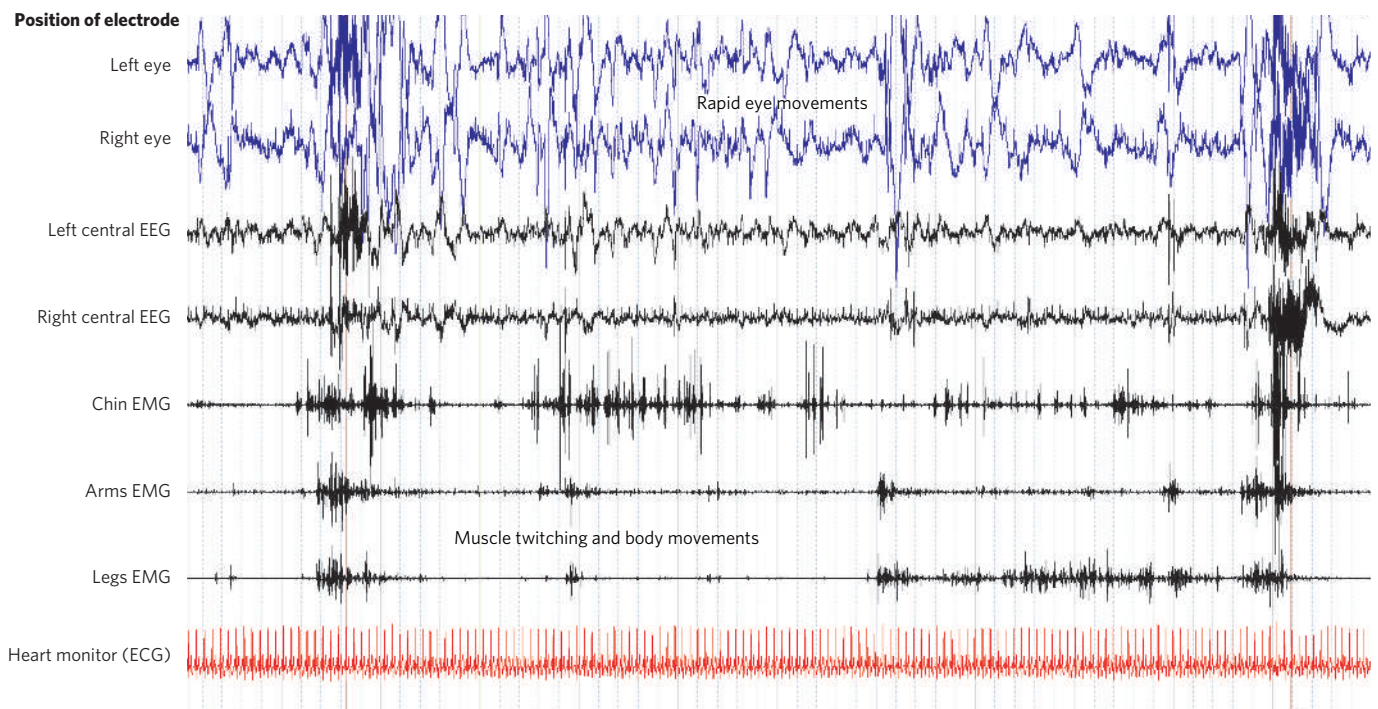


Figure 5 | REM sleep without atonia in a patient with REM sleep behaviour disorder. Note the prominent muscle activity and body movements during REM sleep, which is normally accompanied by active muscle paralysis. EEG, electroencephalogram; EMG, electromyogram.

nature of these episodes. The potential for injury to self or a bed partner raises interesting and difficult forensic medicine issues. Two striking demographic characteristics of RBD are that it predominantly affects males (about 90%) and usually begins after the age of 50 years. Chronic RBD may be preceded by a lengthy prodrome. RBD in humans occurs in both an acute and chronic form. The acute form is often due to undesirable side-effects of prescribed medications, most commonly antidepressant medications, particularly the serotonin-specific reuptake inhibitors (SSRIs)⁶⁰.

The chronic form of RBD is usually either idiopathic or associated with neurological disorders. One of the most interesting implications of chronic RBD is its growing association with neurodegenerative disorders, particularly the synucleinopathies (Parkinson's disease, multiple system atrophy or dementia with Lewy body disease). RBD may be the first manifestation of these conditions, and may precede any other manifestation of the underlying neurodegenerative process by more than 10 years⁶¹. As patients with RBD are followed over time, it is becoming clear that more are developing neurodegenerative disorders. This has led to the suggestion that the term 'idiopathic' RBD be replaced by 'cryptogenic' RBD. The striking relationship between RBD and neurodegenerative disorders has led to fascinating neuroimaging and cerebral blood flow studies in patients with RBD^{62–68}. Such clinical/neuroimaging studies in a variety of neurological conditions will undoubtedly reveal much about brain function and sleep/wakefulness. The fact that RBD may be a very early harbinger of Parkinson's disease will become very important when a neuroprotective pharmacological agent is developed for Parkinson's disease. In keeping with the fact that both narcolepsy and RBD represent state boundary dyscontrol conditions, it should be no surprise that there is a higher incidence of RBD in patients with narcolepsy. Adding to the likelihood that RBD will occur in patients with narcolepsy, medications prescribed to treat cataplexy associated with narcolepsy (tricyclic antidepressants, SSRIs and monoamine oxidase inhibitors) can trigger or exacerbate RBD⁶⁹.

Fortunately, the benzodiazepine clonazepam is a highly effective treatment for RBD⁷⁰. The mechanism of action of clonazepam in RBD is unknown.

Isolated, often bizarre, sleep-related events may be experienced by perfectly normal people, and most do not warrant further extensive or

expensive evaluation. Serious attention should be made to complaints of sleep-related behaviours that are potentially violent or injurious. Forensic issues may result from such behaviours⁵⁷.

Future directions

Progress in basic science sleep physiology and clinical sleep medicine is rapid at the moment. Knowledge gained from sleep research in insects and animals will undoubtedly lead to even more effective treatments for the myriad human sleep disorders. Identification of sleep and circadian abnormalities in transgenic mouse models of Alzheimer's disease and Huntington's disease are just two examples^{71–73}. The evolving relationship between RBD and neurodegenerative disorders and sleep in the prion diseases (Creutzfeldt–Jakob, fatal familial insomnia and bovine encephalopathy) is another. Many aspects of sleep and wakefulness are genetically influenced⁷⁴. Increasingly, studies in *Drosophila* are providing important information regarding genetic influences in sleep patterns (duration and timing of sleep)^{75–77}. In addition, questions raised by human sleep disorders will help direct basic science research. Close collaboration between basic science sleep researchers and sleep medicine clinicians will continue to result in great discovery.

1. National Commission on Sleep Disorders Research. Report of the National Commission on Sleep Disorders Research (Research DHHS Pub. No. 92. Supplier of Documents, U. S. Government Printing Office, Washington, DC, 1992).
2. Caples, S. M., Gamli, A. S. & Somers, V. K. Obstructive sleep apnea. *Ann. Intern. Med.* **142**, 187–197 (2005).
3. Guillemlault, C. & Framherz, S. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 780–790 (Elsevier Saunders, Philadelphia, 2005).
4. Nobili, L. et al. Ultradian aspects of sleep in narcolepsy. *Neurophysiol. Clin.* **26**, 51–59 (1996).
5. Bassetti, C. & Aldrich, M. S. Narcolepsy. *Neurologic Clinics* **14**, 545–571 (1996).
6. Mignot, E. Genetic and familial aspects of narcolepsy. *Neurology* **50** (suppl. 1), S16–S22 (1998).
7. Taheri, S., Zeitzer, J. M. & Mignot, E. The role of hypocretins (orexins) in sleep regulation and narcolepsy. *Annu. Rev. Neurosci.* **25**, 283–313 (2002).
8. Mignot, E. et al. The role of cerebrospinal fluid hypocretin measurement in the diagnosis of narcolepsy and other hypersomnias. *Arch. Neurol.* **59**, 1553–1562 (2002).
9. Johns, J., Wu, M.-F. & Siegel, J. M. Systemic administration of hypocretin-1 reduces cataplexy and normalizes sleep and waking durations in narcoleptic dogs. *Sleep Res. Online* **3**, 23–28 (2000).
10. Bonnet, M. H. & Arand, D. L. Activity, arousal, and the MSLT in patients with insomnia. *Sleep* **23**, 205–212 (2000).

11. Stepanski, E., Zorick, F., Roehrs, T., Young, D. & Roth, T. Daytime alertness in patients with chronic insomnia compared with asymptomatic control subjects. *Sleep* **11**, 54–60 (1998).
12. Vgontzas, A. N. *et al.* Chronic insomnia is associated with nyctohemeral activation of the hypothalamic-pituitary-adrenal axis: clinical implications. *J. Endocrinol. Metab.* **86**, 3787–3794 (2001).
13. Perlis, M. L., Merica, H., Smith, M. T. & Giles, D. E. Beta EEG activity and insomnia. *Sleep Med. Rev.* **5**, 365–376 (2001).
14. Smith, M. T. *et al.* Neuroimaging of NREM sleep in primary insomnia: a Tc-99-HMPAO single photon emission computed tomographic study. *Sleep* **25**, 325–335 (2002).
15. Lasagna, L. Over-the-counter hypnotics and chronic insomnia in the elderly. *J. Clin. Psychopharmacol.* **15**, 383–386 (1995).
16. Mendelson, W. B. *et al.* The treatment of chronic insomnia: drug indications, chronic use and abuse liability. Summary of a 2001 new clinical drug evaluation unit meeting symposium. *Sleep Med. Rev.* **8**, 7–17 (2004).
17. Schenck, C. H. & Mahowald, M. W. Long-term, nightly benzodiazepine treatment of injurious parasomnias and other disorders of disrupted nocturnal sleep in 170 adults. *Am. J. Med.* **100**, 333–337 (1996).
18. Hajak, G., Bandelow, B., Zulley, J. & Pittrow, D. 'As needed' pharmacotherapy combined with stimulus control treatment in chronic insomnia — assessment of a novel interventional strategy in a primary care setting. *Ann. Clin. Psychiatry* **14**, 1–7 (2002).
19. Morin, C. M., Bastien, C. H., Brink, D. & Brown, T. R. Adverse effects of temazepam in older adults with chronic insomnia. *Hum. Psychopharmacol. Clin. Exper.* **18**, 75–82 (2003).
20. Dawson, D. & van den Heuvel, C. J. Integrating the actions of melatonin on human physiology. *Ann. Med.* **30**, 95–102 (1998).
21. Phillips, B. *et al.* Epidemiology of restless legs symptoms in adults. *Arch. Intern. Med.* **160**, 2137–2141 (2000).
22. Trenkwalder, C. *et al.* Circadian rhythm of periodic limb movements and sensory symptoms of restless legs syndrome. *Mov. Dis.* **14**, 102–110 (1999).
23. Winkelman, J. *et al.* 'Anxietas Tibiarum'. Depression and anxiety disorders in patients with restless legs syndrome. *J. Neurol.* **252**, 67–71 (2005).
24. Desautels, A. *et al.* Restless legs syndrome. Confirmation of linkage to chromosome 12q, genetic heterogeneity, and evidence of complexity. *Arch. Neurol.* **62**, 591–596 (2005).
25. Allen, R. P. & Earley, C. J. Restless legs syndrome: a review of clinical and pathophysiologic features. *J. Clin. Neurophysiol.* **18**, 128–147 (2001).
26. Bucher, S. F., Seelos, K. C., Oertel, W. H., Reiser, M. & Trenkwalder, C. Cerebral generators involved in the pathogenesis of the restless legs syndrome. *Ann. Neurol.* **41**, 639–645 (1997).
27. Turjanski, N., Lees, A. J. & Brooks, D. J. Striatal dopaminergic function in restless legs syndrome. *Neurology* **52**, 932–937 (1999).
28. Tergau, F., Wischer, S. & Paulus, W. Motor system excitability in patients with restless legs syndrome. *Neurology* **52**, 1060–1063 (1999).
29. Stiasny-Kolster, K., Magerl, W., Oertel, W. H., Moller, J. C. & Treede, R.-D. Static mechanical hyperalgesia without dynamic tactile allodynia in patients with restless legs syndrome. *Brain* **127**, 773–782 (2004).
30. von Spiczak, S. *et al.* The role of opioids in restless legs syndrome: an [¹¹C]diprenorphine PET study. *Brain* **128**, 906–917 (2005).
31. Moore-Ede, M. C., Sulzman, F. M. & Fuller, C. A. *The Clocks That Time Us: Physiology of the Circadian Timing System* (Harvard Univ. Press, Cambridge, MA, 1982).
32. Sack, R. L., Lewy, A. J., Blood, M. L., Keith, L. D. & Nakagawa, H. Circadian rhythm abnormalities in totally blind people; incidence and clinical significance. *J. Clin. Endocrinol. Metab.* **75**, 127–134 (1992).
33. Sack, R. L., Lewy, A. J. & Hughes, R. J. Guidelines for prescribing melatonin for sleep and circadian rhythm disorders. *Ann. Med.* **30**, 115–121 (1998).
34. Czeisler, C. A., Richardson, G. S. & Coleman, R. M. Chronotherapy: resetting the circadian clocks of patients with delayed sleep phase insomnia. *Sleep* **4**, 1–21 (1981).
35. Reid, K. J. & Zee, P. C. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 691–701 (Elsevier Saunders, Philadelphia, 2005).
36. Yamadera, H., Takahashi, K. & Okawa, M. A multicenter study of sleep-wake rhythm disorders: clinical features of sleep-wake rhythm disorders. *Psychiatr. Clin. Neurosci.* **50**, 195–201 (1996).
37. Xu, W. *et al.* Functional consequences of a CKI-delta mutation causing familial advanced sleep phase syndrome. *Nature* **434**, 640–644 (2005).
38. Toh, K. L. *et al.* An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* **291**, 1040–1043 (2001).
39. Archer, S. N. *et al.* A length polymorphism in the circadian clock gene Per3 is linked to delayed sleep phase syndrome and extreme diurnal preference. *Sleep* **26**, 413–415 (2003).
40. Dagan, Y., Yovel, I., Hallis, D., Eisenstein, M. & Raichik, I. Evaluating the role of melatonin in the long-term treatment of delayed sleep phase syndrome (DSPS). *Chronobiol. Inter.* **15**, 181–190 (1998).
41. Mahowald, M. W. & Schenck, C. H. Evolving concepts of human state dissociation. *Arch. Ital. Biol.* **139**, 269–300 (2001).
42. Mahowald, M. W. & Schenck, C. H. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 786–795 (W. B. Saunders, Philadelphia, 2000).
43. Fisher, C., Kahn, E., Edwards, A. & Davis, D. M. A psychophysiological study of nightmares and night terrors. I. Physiological aspects of the stage 4 night terror. *J. Nerv. Ment. Dis.* **157**, 75–98 (1973).
44. Fisher, C., Kahn, E., Edwards, A., Davis, D. M. & Fine, J. A psychophysiological study of nightmares and night terrors. III. Mental content and recall of stage 4 night terrors. *J. Nerv. Ment. Dis.* **158**, 174–188 (1974).
45. Schenck, C. H., Hurwitz, T. D., Bundlie, S. R. & Mahowald, M. W. Sleep-related injury in 100 adult patients: a polysomnographic and clinical report. *Am. J. Psychiatry* **146**, 1166–1173 (1989).
46. Guilleminault, C., Moscovitch, A. & Leger, D. Forensic sleep medicine: nocturnal wandering and violence. *Sleep* **18**, 740–748 (1995).
47. Llorente, M. D., Currier, M. B., Norman, S. & Mellman, T. A. Night terrors in adults: phenomenology and relationship to psychopathology. *J. Clin. Psychiatry* **53**, 392–394 (1992).
48. Shapiro, C. M., Trajanovic, N. N. & Fedoroff, J. P. Sexomnia — a new parasomnia? *Can. J. Psychiatry* **48**, 311–317 (2003).
49. Winkelman, J. W., Herzog, D. B. & Fava, M. The prevalence of sleep-related eating disorder in psychiatric and non-psychiatric populations. *Psychol. Med.* **29**, 1461–1466 (1999).
50. Schenck, C. H. & Mahowald, M. W. Review of nocturnal sleep-related eating disorders. *Int. J. Eat. Dis.* **15**, 343–356 (1994).
51. Rosenfeld, D. S. & Elhajjar, A. J. Sleepsex: a variant of sleepwalking. *Arch. Sex. Behav.* **27**, 269–278 (1998).
52. Rosen, G., Mahowald, M. W. & Ferber, R. in *Principles and Practice of Sleep Medicine in the Child* (eds Ferber, R. & Kryger, M.) 99–106 (W. B. Saunders, Philadelphia, 1995).
53. Nino-Murcia, G. & Dement, W. C. in *Psychopharmacology: the Third Generation of Progress* (ed. Meltzer, H. Y.) 873–879 (Raven, New York, 1987).
54. Ohayon, M., Guilleminault, C., Priest, R. G. Night terrors, sleepwalking, and confusional arousals in the general population: their frequency and relationship to other sleep and mental disorders. *J. Clin. Psychiatry* **60**, 268–276 (1999).
55. Klackenberg, G. Somnambulism in childhood — prevalence, course and behavior correlates. A prospective longitudinal study (6–16 years). *Acta Paediatr. Scand.* **71**, 495–499 (1982).
56. Kahn, E., Fisher, C. & Edwards, A. in *The Mind in Sleep. Psychology and Psychophysiology* (eds Ellman, S. D. & Antrobus, J. S.) 437–447 (John Wiley & Sons, New York, 1991).
57. Mahowald, M. W. & Schenck, C. H. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 960–968 (Elsevier/Saunders, Philadelphia, 2005).
58. Mahowald, M. W. & Cramer-Bornemann, M. A. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 889–896 (Elsevier/Saunders, Philadelphia, 2005).
59. Schenck, C. H. in *Encyclopedia of the Neurological Sciences* (eds Aminoff, M. J. & Daroff, R. B.) 146–149 (Academic, San Diego, 2003).
60. Mahowald, M. W. & Schenck, C. H. in *Principles and Practice of Sleep Medicine* (eds Kryger, M. H., Roth, T. & Dement, W. C.) 897–916 (Elsevier/Saunders, Philadelphia, 2005).
61. Boeve, B. F. *et al.* Synucleinopathy pathology often underlies REM sleep behavior disorder and dementia or parkinsonism. *Neurology* **61**, 40–45 (2003).
62. Eiseensehr, I. *et al.* Reduced striatal dopamine transporters in idiopathic rapid eye movement sleep behavior disorder. Comparison with Parkinson's disease and controls. *Brain* **123**, 1155–1160 (2000).
63. Eiseensehr, I. *et al.* Increased muscle activity during rapid eye movement sleep correlates with decrease of striatal presynaptic dopamine transporters. IPT and IBZM SPECT imaging in subclinical and clinically manifest idiopathic REM sleep behavior disorder, Parkinson's disease, and controls. *Sleep* **26**, 507–512 (2003).
64. Albin, R. L. *et al.* Decreased striatal dopaminergic innervation in REM sleep behavior disorder. *Neurology* **55**, 1410–1412 (2000).
65. Shirakawa, S.-I. *et al.* Study of image findings in rapid eye movement sleep behavioral disorder. *Psychiatr. Clin. Neurosci.* **56**, 291–292 (2002).
66. Fantini, M. L. *et al.* Slowing of electroencephalogram in rapid eye movement sleep behavior disorder. *Ann. Neurol.* **53**, 774–780 (2003).
67. Miyamoto, M. *et al.* Brainstem function in rapid eye movement sleep behavior disorder: the evaluation of brainstem function by proton MR spectroscopy (1H-MRS). *Psychiatr. Clin. Neurosci.* **54**, 350–351 (2000).
68. Gilman, S. *et al.* REM sleep behavior disorder is related to striatal monoaminergic deficit in MSA. *Neurology* **61**, 29–34 (2003).
69. Schenck, C. H. & Mahowald, M. W. Motor dyscontrol in narcolepsy: rapid-eye-movement (REM) sleep without atonia and REM sleep behavior disorder. *Ann. Neurol.* **32**, 3–10 (1992).
70. Schenck, C. H. & Mahowald, M. W. Polysomnographic, neurologic, psychiatric, and clinical outcome report on 70 consecutive cases with REM sleep behavior disorder (RBD): sustained clonazepam efficacy in 89.5% of 57 treated patients. *Cleveland Clin. J. Med.* **57** (Suppl.), S9–S23 (1990).
71. Petersen, A. *et al.* Orexin loss in Huntington's disease. *Hum. Mol. Genet.* **14**, 39–47 (2005).
72. Morton, A. J. *et al.* Disintegration of the sleep-wake cycle and circadian timing in Huntington's disease. *J. Neurosci.* **25**, 1573–163 (2005).
73. Wisor, J. P. *et al.* Sleep and circadian abnormalities in a transgenic mouse model of Alzheimer's disease: a role for cholinergic transmission. *Neuroscience* **131**, 375–385 (2005).
74. Dauvilliers, Y., Maret, S. & Tafti, M. Genetics of normal and pathological sleep. *Sleep Med. Rev.* **9**, 91–100 (2005).
75. Huber, R. *et al.* Sleep homeostasis in *Drosophila melanogaster*. *Sleep* **27**, 628–639 (2004).
76. Hendricks, J. C. & Sehgal, A. Why a fly? Using *Drosophila* to understand the genetics of circadian rhythms and sleep. *Sleep* **27**, 334–342 (2004).
77. Cirelli, C. *et al.* Reduced sleep in *Drosophila Shaker* mutants. *Nature* **434**, 1087–1092 (2005).

Supplementary Information A video sequence is available linked to the online version of the paper at www.nature.com/nature.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence should be addressed to M.W.M. (mahow002@umn.edu).

What are the memory sources of dreaming?

Tore A. Nielsen^{1,2} and Philippe Stenstrom^{1,3}

Investigators since Freud have appreciated that memories of the people, places, activities and emotions of daily life are reflected in dreams but are typically so fragmented that their predictability is nil. The mechanisms that translate such memories into dream images remain largely unknown. New research targeting relationships between dreaming, memory and the hippocampus is producing a new theory to explain how, why and when we dream of waking life events.

The night-time production of dreams is an unexplained marvel of human existence. The function of dreams and their underlying brain mechanisms continue to be matters of intense academic debate. In this article, we consider several new avenues of research that may help to overcome long-standing obstacles to explaining the functions and mechanisms of dreaming. The research concerns how episodic memories (memories of personally lived experiences) are consolidated over time and how they may be modified during dreaming (Box 1). They are altered in such a way that their autobiographical origins are obscured even though the subjective context within which they appears is a credible simulation of reality.

Memory sources are a key to dream formation

Although the concept of episodic memory was unknown to Freud, he considered such memories to be instrumental in the formation of dreams¹. He coined the term 'day-residue' to refer to elements that connected with experiences of the previous day and that he identified in all of the dreams he scrutinized. His detailed descriptions of day-residues in his patients' dreams as well as his theoretical model describing their transformation from memories into dream elements became central to psychoanalytic theory and therapy. These elaborate 'dreamwork mechanisms' remain largely untested today.

When psychophysiological methods were introduced into the study of dreaming, the pursuit of memory sources continued with the aid of polysomnographic identification of rapid eye movement (REM) and non-REM (NREM) sleep stages and a variety of experimental methods for tracking these sources. Methods such as pre-sleep stimulation (for example, provocative films and virtual environments), sensory stimulation during sleep (for example, tones, shocks and odours) and subjects' post-hoc identification of memory sources of dream elements²⁻⁴ demonstrated the penetrability of dreaming to experimental and naturally occurring episodic stimuli. Results confirmed the robustness of day-residue memory elements and largely attributed these elements to the fragmentation and transformation of episodic memories. More recent studies⁵ suggest that dreams only rarely portray complete episodic memories, that is, reproductions of ensembles of places, actions and characters; such memories occur in a mere 1.4% of reports. Isolated fragments or features of episodic memories are more common, however. The reproduction of isolated spatial or temporal features of memories occurred in 28–38% of reports in one study³; 65% of dream elements were linked to features of waking events in another study⁵.

In very emotional dreams such as nightmares caused by traumatic experiences, dream imagery can become highly episodic. The traumatic event may be replayed veridically as an ensemble of coherent

spatio-temporal, perceptual and emotional details. It remains to be determined whether the auto-noetic (or 'self-in-time') awareness of episodic memories (Box 1) is also preserved during such dreams.

If dreaming is not reliably episodic as many nightmares are, it nonetheless does simulate reality in several striking respects. Dreams seem to take place in real, spatially coherent, environments with which the self interacts perceptually, for example, by orienting, seeking and assimilating sensory information, much as it does with the real world. The self also seems to engage realistic characters in emotional and intellectual exchanges. Semantic information and a sense of knowing are often also present.

This apparent contradiction between dreaming's lack of a fully episodic structure and its portrayal of coherent virtual worlds raises several theoretical questions. Why do memories appear in dreams as fragments or partial episodes but also occasionally as complete replays? If episodic memory is inactive during REM sleep, as some researchers propose, why does REM sleep deprivation interfere with the consolidation of episodic memories in some cases but not others (see ref. 6 for a review)? Does the partial dissolution of memory episodes in dreaming imply that the basic 'units' of the episodic memory system can be identified from careful study of dream reports? Recent advances in our understanding of the memory functions of sleep (for example, see the review in this issue by Stickgold, p. 1272) are beginning to provide answers. And the study of dreaming is beginning to emerge as one of the most promising new methods for sampling the memory processes concealed beneath the veil of sleep.

The hippocampus and episodic memories in dreaming

A consensus among theorists is emerging that altered hippocampal function during sleep accounts, at least in part, for the absence of complete episodic memories during dreaming. Although they propose different mechanisms of hippocampal action (Table 1), all agree that hippocampal changes contribute to the unique characteristics of dream content. Human brain imaging findings support these speculations. Activity in the hippocampus, entorhinal cortex and other parahippocampal regions is increased during REM sleep relative to both waking and NREM sleep (see ref. 7 for a review) and is correlated with REM sleep eye movements⁸. Rhythmic slow activity (1.5–3.0 Hz) in the parahippocampal formation has also been recorded during REM sleep in human subjects⁹. There is thus ample evidence that the hippocampus is active during REM sleep even though this activity has not yet been linked specifically to changes in episodic memory organization. Just how might such a link be forged?

Three lines of inquiry are, in our view, especially pertinent to this question. Each line of inquiry addresses a particular formal characteris-

¹Dream and Nightmare Laboratory, ²Department of Psychiatry, ³Department of Psychology, Hôpital du Sacré-Cœur de Montréal, 5400 Blvd Gouin Ouest, Montréal, Québec, Canada H4J 1C5.

tic of dream imagery and explores the role of the hippocampus and related structures in the production of this characteristic. Each also points to methodological improvements to the study of dream experience that might shed new light on the nature and consolidation of episodic memory during sleep or wakefulness.

The 'here' and 'now' of dream experience

The first line of inquiry concerns the fact that subjective dream experience is, for the most part, spatially and temporally coherent. Although many characterize this coherence as narrative or story-like continuity, a deeper implication is that a type of spatial and temporal binding underlies dreaming that is analogous to the perceptual binding thought to underlie waking consciousness. Dream-related binding would sustain the moment-to-moment illusion that dreaming is taking place from the first-person perspective ('here') and in the subjective present ('now'). This 'here-and-now' level of coherence would proceed in parallel with other, more global, mechanisms, such as narrative organization, and would probably operate on the threshold of subjective awareness, where it would be accessible for scrutiny by self-observational methods (for example, see ref. 10).

The role of binding disparate spatial and temporal memory elements has been attributed specifically to the hippocampus on the basis of functional magnetic resonance imaging (fMRI) studies (for example, see ref. 11). The hippocampus processes both temporal patterns (for example, temporal ordering of events and 'chunking' of events into perceptible units) and spatial patterns (for example, rapid acquisition of novel configurations and pattern completion)¹², which have been linked to the CA1, CA3 and presubiculum regions of the hippocampus. The latter contain cells that in animals seem to subserve a sense of 'place', 'head-direction' and 'direction-by-place'^{12,13}. All these features could also contribute to the virtual navigation and spatio-temporal realism of dreaming. It may therefore be useful to examine whether such attributes are preserved during dreaming sequences and whether they are tied to hippocampal functioning. Animal studies already suggest that the neuronal correlates of long behavioral sequences are replayed by hippocampal neurons in real-time during REM sleep¹⁴.

More detailed scrutiny of dream experience is clearly needed to clarify such a role for the hippocampus in maintaining the stream of perception-like imagery during sleep. Subjects could be trained to provide more accurate and detailed subjective reports of dreaming at this microstructural level. It may also be necessary to evaluate the contributions of a variety of cognitive systems and their dependence upon hippocampal activity. For example, inference processes, which are highly dependent upon the hippocampus in humans¹⁵, may participate in resolving disparities between dream elements for which the binding process has been unsuccessful.

Studies of the dreams of patients with hippocampal damage, whose memory for recent episodic events is poor, might also shed light on the process of imagery binding. Studies are sparse, but provide a few tantalizing clues to the role of the hippocampus. Torda reports that the dreams of three such patients are infrequent but, when present, are short, stereotyped, repetitious, unemotional and lacking day-residues or symbolic elaborations¹⁶. In fact, they frequently replay actual events, that is, complete episodic memories. Such findings closely parallel the finding that patients with post-traumatic stress disorder suffer both diminished hippocampal capacity and episodic replay nightmares. This is consistent with the notion that an intact hippocampus is essential for access to complete episodic memories during wakefulness and dismantling of them for dream formation during sleep.

Dream sources are masked by chronobiological factors

A second line of inquiry concerns the fact that the memory sources of dreaming are often governed by temporal mechanisms that may obscure their episodic origins. Chronobiological factors at several levels seem to influence the selection of memory sources (for a review, see ref. 17). Oscillatory clocks, with a recurrent periodicity, affect both the selection of episodic memory sources (on a 90-minute REM/NREM cycle) and the selection of dream elements from recent versus remote

time periods (on a circadian, early versus late-night, cycle). Interval timers, with a fixed duration, influence the selection of dream memory sources arising about 12 hours before the dream (day-residues) and those arising about a week before the dream (see ref. 18 for a review; Fig. 1). Even longer interval delays have also been observed. There is a consistent appearance of memory sources that arise from between ages 10 and 19 (ref. 4).

Some of the delayed memory elements in dreams may be explained by a widely held hippocampal model of time-limited memory consolidation¹⁹. It stipulates that the dependence of newly acquired memories on the hippocampus decreases over time whereas their dependence on neocortical structures, such as the medial prefrontal cortex, increases in a complementary fashion. Memories are, in a sense, relocated over time from the hippocampus to the neocortex, where they may be reiterated, for as yet unknown reasons and in qualitatively different forms, in dream content. We have demonstrated qualitative differences in the memory sources of dreams that arise from recent (day-residue) and delayed (about one-week) time periods¹⁸.

The duration of this hippocampus-to-neocortex relocation of memories probably varies for different types or attributes of memories, extending to years in some cases²⁰. But several animal studies suggest that a major transition takes place over a period of about one week (for example, see ref. 21). Hippocampal cell excitability, in particular, increases after an associative learning task (signalling neural plasticity) and then returns to baseline levels only on the seventh post-train-

Box 1 | Do episodic memories occur in dreams?

As first suggested by Tulving in 1972 (see ref. 38 for a review), episodic memories, which are remembered events experienced in the past, are distinct from semantic memories, which are items of general knowledge about the world and do not refer to any specific autobiographical context. More recently, the definition of episodic memory was expanded to include the subjective appreciation of time. It is associated with awareness, or the capacity to appreciate one's existence as extended throughout time. This 'mental time travel' is distinguished from noetic awareness, or a sense of knowing something to be true, which characterizes semantic memory. Episodic memory can also be distinguished from imagined, but non-autobiographical, virtual experiences that are realistic and appear to unfold in the apparent present, for example, fantasy and dreaming.

Researchers agree that dreaming does not express complete episodic memories. And although auto-noetic awareness in dreams has not yet been examined, it is rare for subjects to report that, while dreaming, they were either aware that events arose in the past or were anticipated to occur in the future. In fact, 62.7% of all dream reports contain no temporal references whatsoever³⁹. Rather, dreaming appears to mimic the flow of waking perceptual experience. Episodic fragments (characters, settings, objects) and patterns (for example, emotion sequences), as well as semantic information, undergo binding over time such that illusions of a first-person perspective and a sense of the continuous present are maintained.

Clinical and brain imaging studies link episodic memory and auto-noetic awareness with activity in several prefrontal brain regions (medial, dorsolateral), visual cortex and medial temporal lobe-including hippocampus^{38,40}. Hippocampal regions are especially implicated when the self-referential quality of the memory task is high⁴⁰. Changes in brain function during REM sleep, especially increased activity in the hippocampal formation and decreased activity in prefrontal regions, are consistent with the view that altered episodic memory functioning linked to these brain regions contributes to the unique quality of dream experience.

Identification of the memory sources of dreaming is a two-step process that depends heavily on subjects' self-reflective abilities. First, subjects must accurately recall and report their dream experiences; second, they must consider individual dream contents and link them with a variety of personal memories. The recall and reporting of dreams after awakenings in the sleep laboratory is generally accepted to produce valid observations. However, subjects are usually given no training in self-observation. Their dream reports may thus lack crucial information about image formation or may unwittingly be modified to improve their narrative coherence or comprehensibility. Similarly, without training subjects may fail to identify memory sources for their dreams older than those from the past few days. Or, they may censor memory sources according to subtle demand characteristics of the experiment.

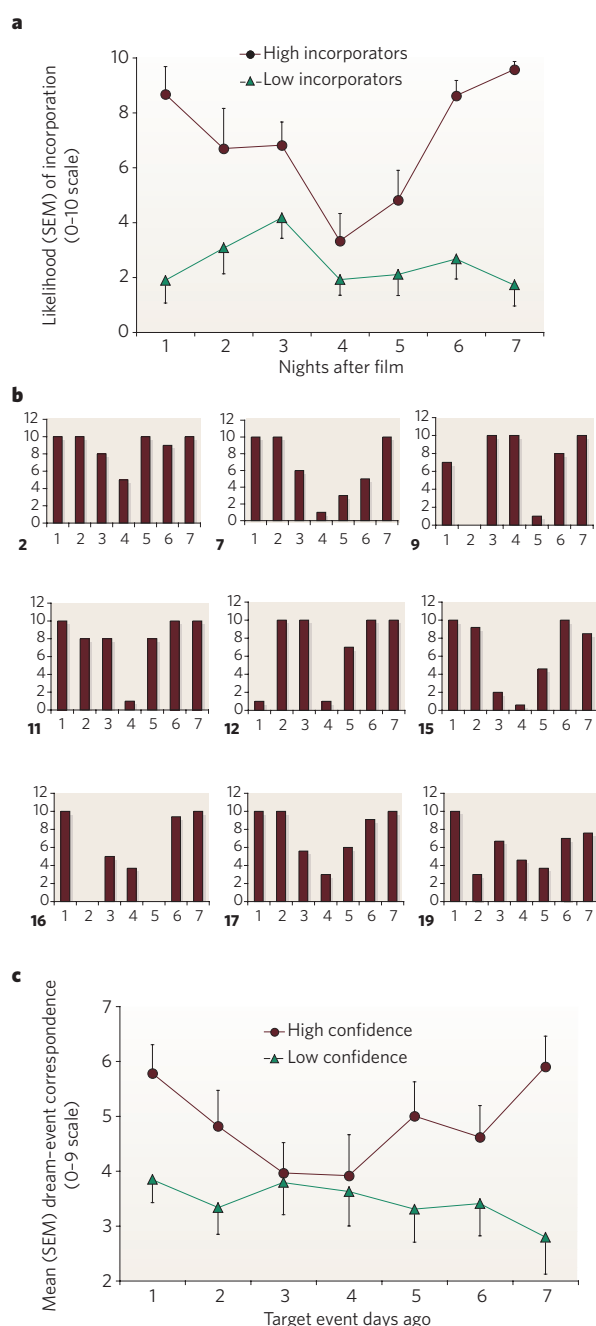


Figure 1 | Temporal variations in access to memory sources. Temporal plots of memory sources of a dramatic event shown in a film appearing in dreams conform to a U-shape. This is consistent with the notion that dreams draw upon memories at different stages of consolidation and from different regions of the brain. A decreasing gradient of access to memories of a given event across the first three nights (recent residues) may correspond to hippocampally mediated memories, whereas an increasing gradient of access to memories from days 5–7 (delayed residues) may correspond to neocortically mediated memories. **a**, Repeated measures design. Exposure to a distressing film about ceremonial sacrifice of a water buffalo produces a U-shaped curve of incorporation likelihood scores across the week for high, but not low, incorporators (data from ref. 41). **b**, Incorporation likelihood scores across the week for nine subjects in the film study who were 'high incorporators', that is, whose dreams were assigned at least one score of 10 on the 0–10 incorporation scale. The U-shaped pattern is apparent in individual subjects' scores. Empty columns indicate that no dream was recalled. **c**, Randomized between-groups design. A U-shaped curve in dream–event correspondence ratings is also observed when subjects randomized to seven different groups search for memory sources from only one of the seven days before their dream and rate their confidence in these sources as high; that is, 5 or greater on a 0–9 scale (data from ref. 18).

ing day²¹. Although no evidence links the temporal attributes of such hippocampal changes directly to delayed memory sources in dreams, three post-learning changes in REM sleep observed in animal models boost the credibility of such a link and suggest fertile avenues for further exploration. First, post-learning REM sleep time is elevated for a period of 5–7 days, as are its accompanying acetylcholine neurotransmitter levels²². Second, task-relevant hippocampal place-cell theta firing during REM sleep reverses polarity over a one-week period, suggesting a weakening of hippocampal synapses²³. Last, partial REM sleep deprivation applied for 15 days after daily maze learning disrupts performance only on days 6–10 (ref. 24).

Delayed memory sources are not typically considered in evaluations of dreaming's episodic memory structure. Our group's experience with one-week-delayed sources is that subjects do not spontaneously report them and must be given specific instructions to identify them. More sophisticated methods of probing subjects' memories, including especially cued autobiographical recall and facilitated self-reflection, and of assessing subjects' confidence in their memory abilities, could be developed to further our understanding of this temporal dimension of episodic memory in relation to dream formation.

Dreams are structured by emotions

A third line of inquiry concerns the observation that in dreams emotional patterns are often preserved, structuring the contents around core relationship patterns²⁵ or an individual's dominant, pre-sleep emotional concerns (for a review see ref. 26).

The emotional sources of dreaming are probably regulated by the amygdala, which controls the encoding and retrieval of emotional memories and the physical expression of emotions. Amygdala activity is higher during REM sleep than during wakefulness⁷ and maintains a reciprocal dependence with the hippocampus in the encoding and storage of memories. For example, hippocampal preprocessing of trace interval durations is essential for the encoding of fear memories in amygdala circuits in mice²⁷. The amygdala also seems capable of gating sensory information through the entorhinal cortex to the hippocampus in rats²⁸. For these reasons, further investigation of amygdalar processes in relation to the hippocampus and dreaming might help to clarify the neural underpinnings of episodic memory organization during sleep.

Surprisingly, the episodic origins of emotional structures during dreaming are often not obvious to naive subjects. Discovering such emotional sources can be a major component of the insight gained by psychotherapy clients²⁹. Emotional sources may go unnoticed because they are expressed only metaphorically, possibly in the service of building adaptive new contexts to assimilate current concerns²⁶. Thus, an assault victim who suffers frequently from feelings of fear mounting to panic may not notice the resemblance between such feelings and emotions felt during his recurrent nightmares of a growing storm that culminates in a tsunami and shipwreck. Elucidation of the emotional sources of dreaming may require development of methods that sensitize subjects to this subtle but influential level of dream experience while simultaneously controlling for the demand characteristics that may influence self-observation.

Conclusion

Closer scrutiny of the episodic origins of dream content are opening new research vistas on the memory functions of sleep. Although several theorists now hypothesize a role for dreaming in memory consolidation (Table 1), a causal link between dreaming and memory remains to be demonstrated. One hypothesis is that the appearance of any memory elements in dreams facilitates learning simply by reactivating those elements to their original (perception-like) state. Another is that the binding of different elements, perhaps around emotionally relevant themes, strengthens and consolidates those elements³⁰. A third is that dream-related memory consolidation is regulated by oscillatory or interval timers of different frequencies, in a manner analogous to the hippocampally mediated waves of *zif-268* gene upregulation that occur during successive REM sleep episodes after

Table 1 | Recent theories implicating altered hippocampal functioning in REM sleep to absence of complete episodic memories during dreaming

Authors	Hypothesized mechanism	Consequence for dream formation	Possible function
Stickgold, Hobson, Fosse & Fosse ³⁵	In REM sleep, reduced information flow from hippocampus to neocortex; flow of weakly associated semantic contents from neocortex to hippocampus	Illogical sequences of barely related objects, characters and locations lacking spatio-temporal coherence	Selective strengthening of associations and consolidation of semantic memories
Johnson ³⁶	Predominance of theta activity during REM sleep causes neocortex to present incongruous and unpredictable information to hippocampus	'Context memories': composites of elements merged from many separate episodic experiences	Context enables retrieval of episodic memories during other ('non-theta' mode) states, for example, waking
Payne & Nadel ³⁷	Increasing cortisol late in sleep diminishes hippocampal-to-neocortical communication; activates memory fragments without spatio-temporal contexts afforded by hippocampus	Fragmented, bizarre imagery synthesized and 'smoothed' into narrative themes	Strengthening, integration and maintenance of memories
Paller & Voss ³⁰	Sleep strengthens connections among dispersed cortical networks and hippocampal-neocortical connections	Narrative created from fragmentary information	Declarative memories consolidated adaptively, that is, in relation to problems, goals, recent experiences

the induction of long-term potentiation in awake rats³¹.

A final, more widely recognized, hypothesis is that dreaming about newly learned material enhances subsequent recall of that material. Three studies in human subjects support this hypothesis and suggest methodological avenues for future exploration. The first is that inter-related dream elements are better recalled in a surprise morning recall task than are unrelated elements³². The second is that pre-sleep stories are better recalled in the morning when subjects dream frequently about constituents of the stories³³. The third is that completion of a mirror-tracing task leads to reports of dreams that metaphorically represent the task (for example, 'trying to stay on a road')³⁴.

A continuing obstacle to proving whether dreaming plays a causal role in memory consolidation is the difficulty subjects encounter in identifying the source memories of their dreams. Even though the assessment of feature binding, temporal factors and emotional structures all suggest novel methods for eliciting and assessing subject reports, future progress may ultimately depend upon whether subjects can be taught to identify and report the non-obvious minutiae of their dreams and memories with sufficient accuracy that the dreamwork mechanisms connecting the two can be discerned. In an era of high-resolution brain imaging, similarly high-resolution reports of dream imagery may be needed. To achieve this, the method of self-observation preferred by Freud, William James and others may yet prove to be among the most productive — particularly in a domain for which the object of study remains directly observable only by dreamers themselves. ■

- Freud, S. *The Interpretation of Dreams* (Basic Books, New York, 1900).
- Arkin, A. M. & Antrobus, J. S. in *The Mind in Sleep* (eds Ellman, S. J. & Antrobus, J. S.) 265–307 (John Wiley & Sons, New York, 1991).
- Cavallero, C., Foulkes, D., Hollifield, M. & Terry, R. Memory sources of REM and NREM dreams. *Sleep* **13**, 449–455 (1990).
- Grenier, J. et al. Temporal references in dreams and autobiographical memory. *Mem. Cogn.* **33**, 280–288 (2005).
- Fosse, M. J., Fosse, R., Hobson, J. A. & Stickgold, R. J. Dreaming and episodic memory: a functional dissociation? *J. Cogn. Neurosci.* **15**, 1–9 (2003).
- Rauch, G., Desgranges, B., Foret, J. & Eustache, F. The relationships between memory systems and sleep stages. *J. Sleep Res.* **14**, 123–140 (2005).
- Hobson, J. A., Pace-Schott, E. F., Stickgold, R. & Kahn, D. To dream or not to dream — relevant data from new neuroimaging and electrophysiological studies. *Curr. Opin. Neurobiol.* **8**, 239–244 (1998).
- Braun, A. R. et al. Dissociated pattern of activity in visual cortices and their projections during human rapid eye movement sleep. *Science* **279**, 91–95 (1998).
- Bodizs, R. et al. Rhythmic hippocampal slow oscillation characterizes REM sleep in humans. *Hippocampus* **11**, 747–753 (2001).
- Nielsen, T. A. A self-observational study of spontaneous hypnagogic imagery using the upright napping procedure. *Imag. Cogn. Person.* **11**, 353–366 (1992).
- Luo, J. & Niki, K. Does hippocampus associate discontinuous events? Evidence from event-related fMRI. *Hippocampus* **15**, 141–148 (2005).
- Kesner, R. P., Lee, I. & Gilbert, P. A behavioral assessment of hippocampal function based on a subregional analysis. *Rev. Neurosci.* **15**, 333–351 (2004).
- Cacucci, F., Lever, C., Wills, T. J., Burgess, N. & O'Keefe, J. Theta-modulated place-by-direction cells in the hippocampal formation in the rat. *J. Neurosci.* **24**, 8265–8277 (2004).
- Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145–156 (2001).
- Heckers, S., Zalesak, M., Weiss, A. P., Diltman, T. & Titone, D. Hippocampal activation during transitive inference in humans. *Hippocampus* **14**, 153–162 (2004).
- Torda, C. Dreams of subjects with bilateral hippocampal lesions. *Acta Psychiatr. Scand.* **45**, 277–288 (1969).
- Nielsen, T. A. Chronobiological features of dream production. *Sleep Med. Rev.* **8**, 403–424 (2004).
- Nielsen, T. A., Kuiken, D., Alain, G., Stenstrom, P. & Powell, R. Immediate and delayed incorporations of events into dreams: further replication and implications for dream function. *J. Sleep Res.* **13**, 327–336 (2004).
- Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
- Milner, B., Squire, L. R. & Kandel, E. R. Cognitive neuroscience and the study of memory. *Neuron* **20**, 445–468 (1998).
- Thompson, L. T., Moyer, J. R. Jr & Disterhoft, J. F. Transient changes in excitability of rabbit CA3 neurons with a time course appropriate to support memory consolidation. *J. Neurophysiol.* **76**, 1836–1849 (1996).
- Smith, C. *Sleep and Brain Plasticity* (eds Maquet, P., Smith, C. & Stickgold, R.) 117–133 (Oxford Univ. Press, New York, 2003).
- Poe, G. R., Nitz, D. A., McNaughton, B. L. & Barnes, C. A. Experience-dependent phase-reversal of hippocampal neuron firing during REM sleep. *Brain Res.* **855**, 176–180 (2000).
- Bjorness, T. E., Riley, B. T., Tysor, M. K. & Poe, G. R. REM restriction persistently alters strategy used to solve a spatial task. *Learn. Mem.* **12**, 352–359 (2005).
- Popp, C. A. et al. Repetitive relationship themes in waking narratives and dreams. *J. Consult. Clin. Psychol.* **64**, 1073–1078 (1996).
- Hartmann, E. *Dreams and Nightmares: the New Theory on the Origin and Meaning of Dreams*. (Plenum, New York, 1998).
- Misane, I. et al. Time-dependent involvement of the dorsal hippocampus in trace fear conditioning in mice. *Hippocampus* **15**, 418–426 (2005).
- Kajiwara, R., Takashima, I., Mimura, Y., Witter, M. P. & Iijima, T. Amygdala input promotes spread of excitatory neural activity from perirhinal cortex to the entorhinal-hippocampal circuit. *J. Neurophysiol.* **89**, 2176–2184 (2003).
- Pesant, N. & Zadra, A. Working with dreams in therapy: what do we know and what should we do? *Clin. Psychol. Rev.* **24**, 489–512 (2004).
- Paller, K. A. & Voss, J. L. Memory reactivation and consolidation during sleep. *Learn. Mem.* **11**, 664–670 (2004).
- Ribeiro, S. et al. Induction of hippocampal long-term potentiation during waking leads to increased extrahippocampal *zif-268* expression during ensuing rapid-eye-movement sleep. *J. Neurosci.* **22**, 10914–10923 (2002).
- Cipolli, C., Fagioli, I., Mazzetti, M. & Tuozzi, G. Consolidation effect of the processing of declarative knowledge during human sleep: evidence from long-term retention of interrelated contents of mental sleep experiences. *Brain Res. Bull.* **65**, 97–104 (2005).
- Fiss, H., Kremer, E. & Lichtman, J. The mnemonic function of dreaming. *Sleep Res.* **6**, 122–136 (1977).
- Smith, C. & Hanke, J. Memory processing reflected in dreams from rapid eye movement sleep. *Sleep* **27** (Suppl. 1), A60 (2004).
- Stickgold, R., Hobson, J. A., Fosse, R. & Fosse, M. Sleep, learning, and dreams: off-line memory reprocessing. *Science* **294**, 1052–1057 (2001).
- Johnson, J. D. REM sleep and the development of context memory. *Med. Hypotheses* **64**, 499–504 (2005).
- Payne, J. D. & Nadel, L. Sleep, dreams, and memory consolidation: the role of the stress hormone cortisol. *Learn. Mem.* **11**, 671–678 (2004).
- Tulving, E. Episodic memory: from mind to brain. *Annu. Rev. Psychol.* **53**, 1–25 (2002).
- Hall, C. & Van de Castle, R. I. *The Content Analysis of Dreams* (Appleton-Century-Crofts, New York, 1966).
- Cabeza, R. et al. Brain activity during episodic retrieval of autobiographical and laboratory events: an fMRI study using a novel photo paradigm. *J. Cogn. Neurosci.* **16**, 1583–1594 (2004).
- Powell, R. A., Nielsen, T. A., Cheung, J. S. & Cervenka, T. M. Temporal delays in incorporation of events into dreams. *Percept. Mot. Skills* **81**, 95–104 (1995).

Acknowledgements Thanks are due to D. Petit, E. Martel and T. Paquette for editorial assistance. The work was supported by grants to Nielsen from the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to T.A.N. (tore.nielsen@umontreal.ca).

Evolution of indirect reciprocity

Martin A. Nowak¹ & Karl Sigmund^{2,3}

Natural selection is conventionally assumed to favour the strong and selfish who maximize their own resources at the expense of others. But many biological systems, and especially human societies, are organized around altruistic, cooperative interactions. How can natural selection promote unselfish behaviour? Various mechanisms have been proposed, and a rich analysis of indirect reciprocity has recently emerged: I help you and somebody else helps me. The evolution of cooperation by indirect reciprocity leads to reputation building, morality judgement and complex social interactions with ever-increasing cognitive demands.

Humans are the champions of reciprocity. Experiments and everyday experience alike show that what Adam Smith called ‘our instinct to trade, barter and truck’ relies to a considerable extent on the widespread tendency to return helpful and harmful acts in kind. We do so even if these acts have been directed not to us but to others. This has been analysed under the headings of ‘third party altruism’¹ or ‘indirect reciprocity’², and has led to a considerable amount of experimental and theoretical investigation over the past few years.

Direct reciprocity is captured in the principle: ‘You scratch my back, and I’ll scratch yours’. But it is harder to make sense of the principle ‘You scratch my back and I’ll scratch someone else’s’³ or ‘I scratch your back and someone else will scratch mine’ (Fig. 1). Why should this work? Presumably, I will not get my back scratched if it becomes known that I never scratch anybody else’s. Indirect reciprocity, in this view, is based on reputation. But why should anyone care about what I did to a third party?

There are two approaches converging on this issue. One is rooted in social science, the other in evolutionary biology.

The main reason why economists and social scientists are interested in indirect reciprocity is that one-shot interactions between anonymous partners in a global market become increasingly frequent and tend to replace the traditional long-lasting associations and exchanges based on repeated give and take between relatives, neighbours, or members of the same village. A substantial part of our life is spent in the company of strangers⁴, and many transactions are no longer face-to-face. The growth of web-based auctions and other forms of e-commerce is built, to a considerable degree, on reputation and trust^{5–10}. The possibility to exploit such trust raises what economists call moral hazards. How effective is reputation, especially if information is only partial?

In contrast, evolutionary biologists are interested in the emergence of human societies, which constitutes the last (up to now) of the major transitions in evolution¹¹. Unlike other eusocial species, such as bees, ants or termites, humans display a large amount of cooperation between non-relatives^{12–14}. A considerable part of human cooperation is based on moralistic emotions—for instance, anger directed towards cheats, or the proverbial ‘warm inner glow’ felt after performing an altruistic action. Neuro-economic experiments relate these emotions to physiological processes^{15–17}. Intriguingly, humans not only feel strongly about interactions that involve them directly, they also judge the actions between third

parties, as demonstrated by the contents of gossip¹⁸. Indirect reciprocity is therefore likely to be connected with the origins of moral norms. Such norms are evidently to a large extent culture-specific, but the capacity for moral norms seems to be a human universal for which there is little evidence in other species¹⁹.

Because the recent rapid advance of experimental investigations of indirect reciprocity was in large part driven by theory, we shall discuss the modelling approaches before reviewing the experiments. But first we note, in a wider context, that indirect reciprocity seems to require a ‘theory of mind’²⁰. Whereas altruism directed towards kin works because similar genomes reside in different organisms, reciprocal altruism recognizes that similar minds emerge from different brains. It is easy to conceive that an organism experiences as ‘good’ or ‘bad’ anything that affects the organism’s own reproductive fitness in a positive or negative sense. The step from there to judging, as ‘good’ or ‘bad’, actions between third parties, is not obvious. The same terms ‘good’ and ‘bad’ that are applied to pleasure and pain are also used for moral judgements: this linguistic quirk reveals an astonishing degree of empathy, and reflects highly developed faculties for cognition and abstraction.

This review of theoretical and empirical studies of indirect reciprocity stresses the importance of monitoring not only partners in continuing interactions but also all individuals within the social network. Indirect reciprocity requires information storage and transfer as well as strategic thinking and has a pivotal role in the evolution of collaboration and communication. The possibilities for games of manipulation, coalition-building and betrayal are limitless. Indirect reciprocity may have provided the selective challenge driving the cerebral expansion in human evolution.

Direct versus indirect reciprocity

In the terminology based on Hamilton, Trivers and Wilson^{12–14}, an act is said to be altruistic if it is costly to perform but confers a benefit on another individual. In evolutionary biology, costs and benefits are measured in darwinian fitness, which means reproductive success. In other contexts, other utility scales such as monetary rewards may be more appropriate. Reciprocal altruism in its original, ‘direct’ sense is defined as an exchange of altruistic acts between the same two individuals so that, in total, both obtain a net benefit¹. In the simplest model, the altruistic act consists in conferring a benefit b on the recipient at a cost c to the donor. We shall always assume that the cost is smaller than the benefit, so that if the act is returned, both

¹Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA. ²Faculty for Mathematics, University of Vienna, A-1090 Vienna, Austria. ³IIASA, A-2631, Laxenburg, Austria.

individuals experience a gain. The payoff structure yields an instance of the familiar Prisoner's Dilemma game²¹. If both players cooperate, each receives $b - c$, which is better than what they would obtain by both defecting, namely 0. But a unilateral defector would earn b , which is the highest payoff, and the exploited cooperator would pay the cost c without receiving any benefit. The payoff-maximizing move is defecting.

This changes if the game is repeated for several rounds. For simplicity we shall assume that in each round both players decide simultaneously. We could also assume that they alternate, which leads to a slightly different game^{22–24}. The so-called folk theorem on repeated games implies that if the probability for future rounds is sufficiently high, cooperation can be sustained by so-called trigger strategies, which switch to relentless defection as soon as the co-player defects once^{25,26,87}. A rational player must weigh the benefit of exploiting the co-player in one round against the cost of forfeiting collaboration in all future rounds, and would therefore abstain from defection.

In the context of indirect reciprocity, any two players are supposed to interact at most once with each other. Each player can experience many rounds, but never with the same partner twice. Thus it is not possible that a cheat is held to account by the victim. (In a variant of this model, two players could interact on several occasions, one always as the donor, the other as recipient, so that a return is again excluded.) Clearly, trigger strategies can still ensure a cooperative Nash equilibrium, such that if all players use them, no player would have an incentive to deviate. In strategic thinking, only the payoffs matter, not by whom they are provided. In this sense, the step from direct to indirect reciprocity corresponds simply to the step from personal enforcement to community enforcement^{27–30}. However, a trigger strategy prescribing each person to cooperate until the first defection is personally experienced, and thenceforth to defect, hurts the original wrong-doer only after many rounds. A strategy triggered by the first defection in the population leaves cooperation at the mercy of the first wrong move. In both cases many innocents would be punished, and errors would cause havoc. Obviously, retaliation should be directed towards the cheat rather than towards the whole community. This requires more detailed information. Game theory shows that even if information is transmitted only locally and errors

occur occasionally, cooperation can be sustained: there exist strategies such that no rational player has an interest in deviating unilaterally²⁸.

In evolutionary game theory it is not assumed that players are rational but only that successful strategies spread—by being inherited, for instance, or copied through imitation or learning³¹. For direct reciprocity, game theoretical analysis and individual-based simulations have shown that a population of defectors can be invaded by a small cluster of retaliators³² or even by a single retaliator³³. Typically, one considers a well-mixed population in which individuals meet randomly and play a series of Prisoner's Dilemma games with each other. What counts is the total payoff. Retaliators compensate for the loss of being exploited by a defector in the first round with long sequences of altruistic exchange with other retaliators. Once cooperation is established, a complex evolution takes place, which depends on the size of the population, the cost-to-benefit ratio, the average number of rounds and the probability of errors^{32,34,35}.

A similar model of indirect reciprocity assumes that, within a well-mixed population, individuals meet randomly, one in the role of the potential donor and the other as a potential recipient (Fig. 2). Each individual experiences several rounds of this interaction in both roles, but never with the same partner twice. Again, all that counts is the total payoff. A player can follow either an unconditional strategy, such as always to cooperate or always to defect, or else a conditional strategy, which discriminates between the potential recipients on the basis of past interactions. In a simple example, a discriminating player can help the co-player if that co-player's score exceeds a certain threshold. A player's score is 0 at birth, increases whenever that player helps and decreases whenever the player withholds help. Individual-based simulations show that if the cost-to-benefit ratio is sufficiently low, and the amount of information about the co-player's past sufficiently high, cooperation based on discrimination can emerge.

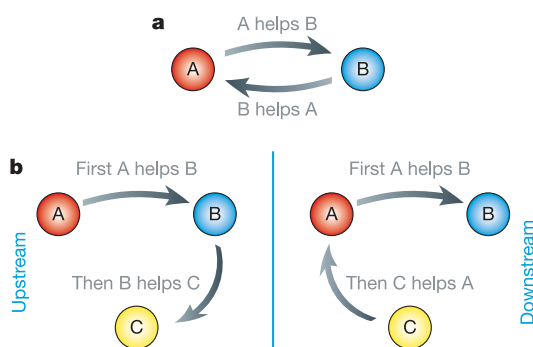


Figure 1 | Direct and indirect reciprocity. **a**, Direct reciprocity means that A helps B and B helps A. **b**, Indirect reciprocity comes in two flavours. 'Upstream reciprocity' (left) is based on a recent positive experience. A person who has been at the receiving end of a donation may feel motivated to donate in turn. Individual B, who has just received help from A, goes on to help C. 'Downstream reciprocity' (right) is built on reputation. Individual A has helped B and therefore receives help from C. Mathematical investigations of indirect reciprocity have shown that natural selection can favour strategies that help others based on their reputation. Upstream reciprocity is harder to understand^{12,56,77,78} but is observed in economic experiments. In both cases, the decision to help can be interpreted as a misdirected act of gratitude. In one case recipients are thanked for what another did; in the other case they are thanked by someone who did not profit by what they did.

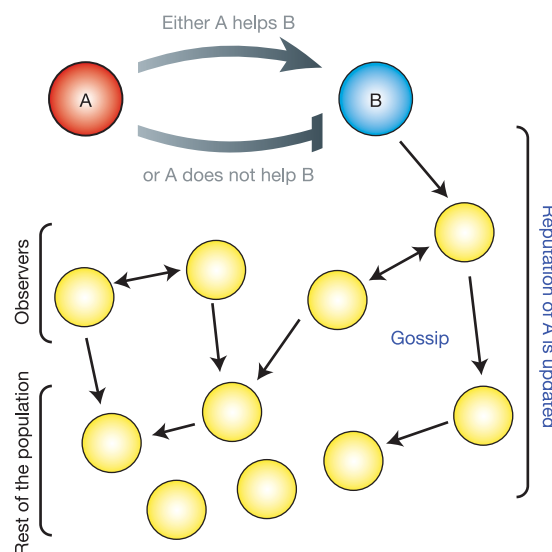


Figure 2 | Building a reputation. In a natural extension of the basic model of indirect reciprocity, an action between donor A and recipient B is observed by a subset of the population³⁶. The observers, the donor and the recipient can inform others. People could pass on what has happened (the action) or their assessment of the action. There are many possibilities of error: the action or the intention of the donor can be interpreted differently by different people; some individuals may receive conflicting information from different sources; some individuals may not receive any information at all; people can have different assessment modules. The reputation of a person is therefore not simply a label that is visible to all others, but instead each person has a private list of the reputation of others. Although language could help to synchronize these lists⁴², ultimately reputation is in the eyes of the beholder.

In the resulting population, help is channelled towards those who have helped^{36–38}.

Two features of this model were immediately apparent^{36,39,40}. One is the paradoxical nature of the discriminating strategy. In terms of rational game theory, why should players care about the scores of others rather than just about their own payoff, and why should they decrease their own score (and thus their likelihood of receiving help on later occasions) by withholding help from low-scorers? Lower scores can cause lower payoffs. The second issue concerns the lack of stability of the cooperative outcome. The simulations display occasional bursts of defection, which are based on a previous build-up of indiscriminating altruists. In a population of discriminators, unconditional cooperators can increase by random drift and eventually invite the invasion of defectors (Fig. 3).

The two issues are closely related. Considered intuitively, the intentions of players who selfishly care about their own income only, and who accordingly give help just to keep their own score high, are vastly different from the intentions of altruists who have only the interests of their co-players in mind and help them on every occasion. But the effect is the same in both cases: support will be given regardless of the co-players' contributions.

Binary assessment, or the world in black and white

To analyse these questions, an even simpler model was proposed, based on a binary score, taking only the values 'good' or 'bad', depending on what the co-player did when last observed^{40–42} (Box 1). This can be viewed as a basic system of moral assessment. In its simplest form (first order), the assessment depends only on the action of the observed player, which means it depends on whether that player gave help or not. A discriminating donor using this assessment rule refuses help to a 'bad' recipient, and therefore becomes 'bad', which reduces the chance of being helped in turn. (With a wider range of score values, a single refusal to help has less impact on the reputation.) Effectively, discriminating players pay a cost for punishing bad co-players. Such a form of altruistic punishment can promote cooperation in the community, but at a

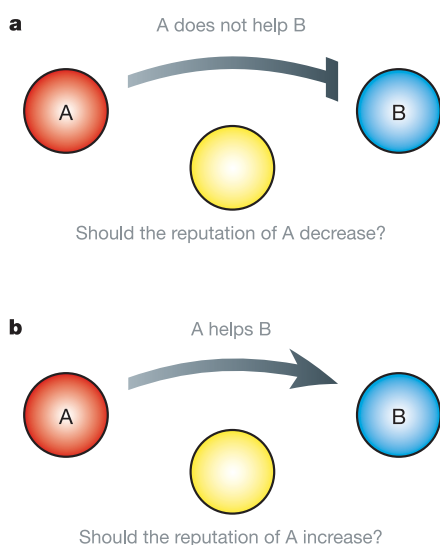


Figure 3 | Two problems with indirect reciprocity. B has defected in previous rounds and therefore has a low reputation. **a**, If A does not help B, so as to punish B for previous defections, then why should the reputation of A be reduced? **b**, If A does help B, although B is a defector, then why should the reputation of A increase? Helping defectors destabilizes cooperation. Strategies (assessment rules) of indirect reciprocity that try to fix these problems are cognitively demanding and vulnerable to deception. Before defecting, A could try to signal that B has a low reputation. We argue that the intricate complexity of indirect reciprocity provided the selective mould for human language and human intelligence.

cost to the punisher, and thus can be viewed as a social dilemma⁴³. The fact that costly punishment has an undisputed role in other contexts, such as public goods games, ultimatum games and trust games^{44–47}, shows that this form of discrimination is plausible.

A more sophisticated assessment rule should distinguish between justified and unjustified defection and should therefore take into account the score of the receiver: someone withholding help from a

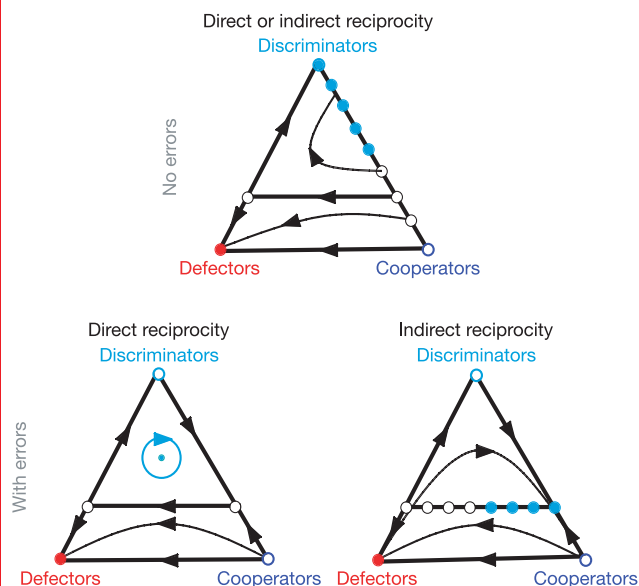
Box 1 | The good, the bad and the discriminating

Similarities and differences between direct and indirect reciprocity become apparent when studying the replicator dynamics of three strategies: always cooperate, always defect and the simplest discriminating strategy. For direct reciprocity, this is Tit For Tat, which helps in the first round and then does whatever the opponent did in the previous round. For indirect reciprocity, this is the strategy that prescribes helping unless the recipient is known to have refused to help in the previous round. In each case, there are many other discriminating strategies, which are likely to take over eventually. This analysis is just a first step.

In the absence of discriminators, defectors win against cooperators. In the absence of cooperators, defectors and discriminators form a bistable system: depending on the initial condition, either one or the other strategy wins. In the absence of defectors, discriminators and cooperators are in equilibrium. Random fluctuations, however, make their frequencies drift up and down. The equilibria can be invaded by defectors if the frequency of discriminators is below a certain threshold. With all three strategies present, the dynamics lead either to defectors only or to a mixture of the two kinds of altruist.

If errors occur, or if an intended donation cannot be implemented through a lack of resources, discriminating and indiscriminating altruists reach, in the absence of defectors, a stable coexistence with a well-defined frequency of discriminators. If all three types are present in the population, the system displays two types of behaviour. If the frequency of discriminators is too low, defectors win. If the discriminators are sufficiently frequent, all three strategies coexist. However, in direct reciprocity the frequencies oscillate periodically, whereas in indirect reciprocity they converge to an equilibrium. In each case, a long series of random fluctuations may eventually destroy the coexistence of the three strategies.

In the deterministic model, only the emergence of other conditional strategies can save cooperation in the long run⁷⁹. For stochastic population dynamics, the time average of the evolutionary oscillations can be centred on discriminators⁸⁰.



Box 1 Figure 1 | Basic evolutionary dynamics of direct and indirect reciprocity.

'bad' player should not pay with a reduced score^{36,39,40,48}. There are many assessment rules of second order, which also depend on the score of the receiver, and of third order, which depend additionally on the score of the donor (Box 2). Only eight of them lead to cooperation and are at the same time evolutionarily stable: a homogeneous population of players using such a strategy cannot be invaded by players using other strategies⁴². All eight of these strategies distinguish justified from unjustified defections. This property must hold for any strategy that maintains cooperation, eliminates cheats and can overcome errors⁴⁹. However, the problem with the concept of justified defection is that it requires information not only about the past of the co-player but also about the past of the co-player's co-players, and their co-players, and so on^{36,40}.

Again, the stability of cooperation is threatened by unconditional altruists who merely wish to keep their own good score. In a population consisting entirely of discriminators, indiscriminating cooperators fare just as well and thus spread by neutral drift. If their frequency exceeds a certain threshold, defectors can invade and take over. The situation is altered in a significant way if players occasionally fail to cooperate even though their strategy prescribes cooperation^{40,50–53}. This can be due to errors in implementation or lack of resources. It is plausible that when recipients need help, donors are also short of means⁵³. In this situation, a population consisting only of conditional and unconditional altruists is not subject to random drift, but selection leads to a well-defined, stable

mix of the two types. This mixture can be vulnerable to the invasion of defectors^{40,41}.

There are several ways out of this impasse. Various assumptions on the distribution of the number of rounds lead to a bistable system in which, depending on the initial state, the population converges either to the fixation of defectors or to a stable mix of altruists that cannot be invaded by defectors^{53,54}. In this case, the very fact that individuals are not perfect and sometimes defect involuntarily promotes the stability of cooperation⁵⁰.

Punishing a player with a bad score creates another player with a bad score; but this 'passing the buck along' can stop in two ways, by encountering either a discriminator who is uninformed or an indiscriminating altruist. Therefore, both the lack of information and the prevalence of unconditional altruists may, surprisingly, stabilize cooperation. A stable mix of discriminating and indiscriminating altruists that cannot be invaded by defectors is also obtained by assuming that, as players grow older, their social network grows and so does their information about their co-players' past^{55,56}. Alternatively, if the discriminating strategy distinguishes between justified and unjustified defection, the population can converge to discriminators only, which cannot be invaded by unconditional strategies⁴⁰.

Can discrimination based on the concept of 'justified defection' be destabilized by errors in perception? Not if players have the same reputation in the eyes of all members of their population. If such a

Box 2 | Let a hundred morals bloom

In a world of binary moral judgements there are four ways of assessing donors in terms of 'first-order assessment': always consider them as good, always consider them as bad, consider them as good if they refuse to give, or consider them as good if they give. Only this last option makes sense. Second-order assessment also depends on the score of the receiver; for example, it can be deemed good to refuse help to a bad person. There are 16 second-order rules. Third-order assessment also depends on the score of the donor; for example, a good person refusing to help a bad person may remain good, but a bad person refusing to help a bad person remains bad. There are 256 third-order assessment rules. We display four of them in Box 2 Fig. 1. With Scoring, cooperation, C, always leads to a good reputation, G, whereas defection, D, always leads to a bad reputation, B. Standing is like Scoring, but it is not bad if a good donor defects against a bad recipient. With Judging, in addition, it is bad to cooperate with a bad recipient. For another assessment rule, Shunning, all donors who meet a bad recipient become bad, regardless of what action they choose. Shunning strikes us as grossly unfair, but it emerges as the winner in a computer tournament if errors in perception are included and if there are only a few rounds in the game⁵⁷.

		Reputation of donor and recipient			
		GG	GB	BG	BB
Action of donor	C	G	G	G	G
	D	B	B	B	B
	C	G	G	G	G
	D	B	B	B	B
		Scoring			
Action of donor	C	G	G	G	G
	D	B	G	B	B
	C	G	B	G	B
	D	B	G	B	B
		Standing			
Action of donor	C	G	B	G	B
	D	B	G	B	B
	C	G	B	G	B
	D	B	G	B	B
		Judging			
Action of donor	C	G	B	G	B
	D	B	B	B	B
	C	G	B	G	B
	D	B	B	B	B
		Shunning			

Reputation of donor after the action

Box 2 Figure 1 | Four assessment rules.

An action rule for indirect reciprocity prescribes giving or not giving, depending on the scores of both donor and recipient. For example, you may decide to help if the recipient's score is good or your own score is bad. Such an action might increase your own score and therefore increase the chance of receiving help in the future. There are 16 action rules.

If we view a strategy as the combination of an action rule and an assessment rule, we obtain 4,096 strategies. In a remarkable calculation, Ohtsuki & Iwasa^{42,49} analysed all 4,096 strategies and proved that only eight of them (the 'leading eight'; Box 2 Fig. 2) are evolutionarily stable under certain conditions and lead to cooperation.

The three asterisks in the assessment module of the 'leading eight' indicate a free choice between G and B. There are therefore $2^3 = 8$ different assessment rules. The action module is built as follows: if the column in the assessment module is G and B, then the corresponding action is C, otherwise the action is D.

Both Standing and Judging belong to the leading eight, but neither Scoring nor Shunning do. However, we expect that Scoring has a similar role in indirect reciprocity to that of Tit For Tat in direct reciprocity. Neither strategy is evolutionarily stable, but their ability to catalyse cooperation in adverse situations and their simplicity constitute their strength. In extended versions of indirect reciprocity, in which donors can sometimes deceive others about the reputation of the recipient, Scoring is the 'foolproof' concept of 'I believe what I see'. Scoring judges the action and ignores the stories.

		GG	GB	BG	BB
Assessment	C	G	*	G	*
	D	B	G	B	*
		C	D	C	C/D
		Action			

Note: if a 'good' donor meets a 'bad' recipient, the donor must defect, and this action does not reduce his reputation.

Box 2 Figure 2 | Ohtsuki & Iwasa's 'leading eight'.

consensual assessment can be achieved, the corresponding strategy is robust^{39,40,42}. But if players have different views about the reputation of others, then errors in perception can undermine cooperation⁵⁷. Such private lists of the scores of co-players are very plausible if individual interactions are observed by only a fraction of the population³⁶. Gossip might be a way of achieving consensus, but it can also be used for spreading unfounded rumours and manipulating co-players. The co-evolution of human language and cooperation by indirect reciprocity is a fascinating and as yet unexplored topic.

Another debated issue concerns the underlying population structure³⁹. Analytical models are often based on the idealization of a very large, well-mixed population. Individual-based simulations typically assume population sizes of 50–100 individuals, on the basis of estimates for the group size of hunter–gatherers. In such small populations, random drift can strongly affect evolution. Populations consisting of many separate groups with a modicum of exchange between them have also been modelled^{39,54}. However, such a population structure facilitates the evolution of altruism through group selection⁵⁷. It could be that cooperators fare less well than defectors within each group but that groups of cooperators fare better than groups of defectors⁵⁸. In extreme cases this leads to cooperation even in the absence of indirect reciprocity. Although it is most interesting to study the interaction between group selection and indirect reciprocity, it is equally important not to confuse the two effects.

Expensive games and economic experiments

The basic experimental set-up for testing indirect reciprocity studies a group of players equipped with an initial monetary endowment. Each player is repeatedly given the opportunity of donating money to a specific co-player, thus increasing the account of this person by an amount b . Players know that if they choose to do so, an amount c will be deducted from their own account—the cost for providing the gift. To eliminate confounding effects, players usually do not interact face-to-face but are given some information about the past actions of their potential recipients. They know that their recipients will never be their donors on future occasions and therefore that there is no scope for direct reciprocity. The interactions both with the co-players and with the experimenters are kept as anonymous as possible,

usually under double-blind conditions. Many parameters can be varied within this basic set-up, for instance the cost-to-benefit ratio, the size of the starting account, the number of interactions, the size of the group, the degree of information about the co-players' behaviour, the length of the game or the social backgrounds of the players (Box 3).

From the first experiments onwards, it was clear that a substantial proportion of the players frequently decide to donate. The propensity for indirect reciprocation is apparently widespread. As expected, donations occur more frequently if the cost-to-benefit ratio is lower or the starting account is higher. Reputation has a considerable influence on the decisions. In particular, the image score of potential recipients correlates well with their expectation to actually receive money⁵⁹. Players who donate less often display a higher degree of discrimination. Players who are more open-handed care less about the recipient's score. Conversely, if players know that their own score is passed on, they are much more likely to donate than otherwise⁶⁰. Many players donate even when they are assured of complete anonymity, possibly because they are not fully convinced. Recent experiments suggest that a nagging suspicion remains: 'What if someone is watching?' Intriguingly, even stylized eyespots suffice to influence the giving behaviour⁶¹.

The hypothesis that more information leads to more cooperation has been confirmed in experiments, which compare three information conditions⁶². In one condition, players have no information about their co-players; in the second they are told about what their co-players have decided when last in the role of a donor; and in the third they also know about the score of the recipient of the co-player. We note that this is not always enough to decide whether a previous defection was justified or not. However, the additional knowledge did enhance cooperation⁶².

In this series of experiments there is a significant positive correlation between the number of gifts given and received, but a slightly negative correlation between the number of gifts given and the total payoff obtained. In another experiment⁶³, however, those who give

Box 3 | Games of cooperation

In the Trust Game there are two players, one in the role of the donor, the other in the role of the responder. The donor can transfer some money to the responder. Upon arrival, the amount is multiplied by three. The responder, then, has the possibility of sending some of it back to the donor. A responder with an income-maximizing strategy should send nothing back. Any donor expecting this should therefore transfer nothing. In real experiments, many donors transfer substantial amounts, and some obtain large returns, so that both players win.

In the Public Goods Game, each of N players can independently decide to transfer some money to a common pool, where it is multiplied by some factor r (smaller than N) and then divided equally between all players irrespective of whether they have contributed or not. Because each player receives, in return for his or her own contribution, only the fraction r/N , the income-maximizing strategy is to contribute nothing. However, in real experiments many players contribute. If all do, they multiply their endowment by the factor r .

The Public Goods Game for $N = 2$ players has the structure of a Prisoner's Dilemma. Two players who cooperate earn more than two players who defect; but a defector cheating on a cooperator earns the highest payoff, and the exploited cooperator earns the lowest. If two players, in a trust game, are simultaneously in the role of the donor, and then simultaneously in the role of the responder, they play two rounds of a Prisoner's Dilemma.

Experimental economists and experimental psychologists have studied these games, and diverse variations, intensively^{81,82}.

Box 4 | Bidding for trust

Trust, 'a lubricant of social life'⁸³, is essential in many types of economic transaction and is also linked to physiological processes⁸⁴. In the Trust Game, donors who trust their responder will expect to gain from transferring money. In contrast, donors in the indirect reciprocity game know that they can expect no direct return, even if their recipient is trustworthy. All they can gain from the transfer is an increased reputation for altruism and trustworthiness.

Game theory shows that cooperation can be sustained in the indirect reciprocity game if each player carries a label²⁸. The strategy prescribes that players who deviate from it have to be punished (by not being helped) for a number of rounds T . A player's label specifies for how many rounds that player has to be punished. If a donor and a receiver meet, the action prescribed by the strategy depends only on their labels, and their labels will be updated depending on the donor's action. No player has an incentive to deviate if all other players adopt this strategy, and the effect of an error will be overcome after T rounds. However, settling on this strategy, for instance on the specific number T , seems to require an institution able to guarantee honest labelling.

Subscribers to eBay auctions are asked to state, after every transaction, whether they were satisfied with their partner or not. Their partner's score can accordingly increase or decrease by one point. The ratings of eBay members, accumulated over 12 months, are public knowledge. This very crude form of assessment seems to suffice for the purpose of reputation-building and seems to be reasonably proof against manipulation. Social history knows many other instances of public scorekeeping: 'Societies have sewn scarlet letters to people's garments, shaved heads, cut off fingers and given medals to signal to strangers some aspect of an individual's past deeds or misdeeds'³⁰. Reputation mechanisms were also important in the emergence of medieval trade⁸⁵.

often end up with the highest payoff, so that there is a strategic advantage to generosity. The discrepancy could be due to the larger number of rounds—the advantage showed up only after a dozen rounds—or to the fact that players were informed not about the recipients' last move only but about their whole history of giving.

Evidence for strategic reputation building is found in many experiments. Donations are more frequent in earlier rounds, when the player's own reputation has a higher impact on the future income. But several experiments show that even players who know that their score will not be communicated show generous behaviour^{62,64}. Such donors cannot be motivated by selfish interests. For many players, however, the propensity to donate more than doubles if they know that their action will be communicated and will therefore affect their own score. The influence of the recipient's score decreases accordingly. Often there is evidence for a dual motivation—players give donations if the recipient's score is high or their own score is low³⁶.

Experiments investigating whether a player's justified defections lowers his or her chance to receive subsequent donations indicate that cognitive problems challenge the donor⁶⁵. Players faced with a full history of all previous rounds take in general a longer time to reach their decision, suggesting that they attempt to take into account not only their recipient's last move but also that of their recipient's recipient. Nonetheless, the statistics of such games with full information look surprisingly similar to those obtained when players know only the score of the recipient. Moreover, players who justifiably refuse to donate to a defector show an increased tendency to provide donations in the following round, as if to make up for that refusal. This indicates that they expect their refusal to lower their score in the co-players' eyes and that they do not rely on the community's understanding.

Many experiments have shown that players who have just received a donation are more likely to give a donation in turn. There is evidence for 'upstream' indirect reciprocity in cyclical networks: as expected, short loops and a high benefit-to-cost ratio favour cooperation⁶⁶. A variant of the Trust Game has two donor–responder pairs, but such that the transfers are criss-cross: the responder of one donor can return money only to the other donor (and does not know the amount transferred by that donor). The return rates turn out to be no lower than if they were addressed to the player's own donor^{67,68}. In experimental situations that are not based on rigid networks, the decisions of donors also tend to mirror their own recent experience^{59,64}. People are nicer to others if third parties have been nicer to them. More generally, it seems that decisions often depend on both

the donor's payoff and the recipient's score, but such strategies have not been analysed so far.

The strategic links of indirect reciprocity

Indirect reciprocity is situated somewhere between direct reciprocity and public goods. On the one hand it is a game between two players only, the donor and the recipient, but on the other hand it has to be played within a larger group.

Richard Alexander claimed that indirect reciprocity originates from direct reciprocity in the presence of interested audiences². A good strategy for the latter is Observer Tit For Tat^{36,69}. Players using this strategy for the repeated Prisoner's Dilemma are following Tit For Tat, except that in the first round they defect if they know that their co-player, in a previous repeated game against another player, has defected. Observer Tit For Tat relies on reputation in the first round and on personal experience in all further rounds against the same co-player. Conversely, experiments show that if several rounds of a Prisoner's Dilemma are appended to an indirect reciprocity game, the display of the previous score increases a player's probability of cooperating with generous players in the first few rounds. After a couple of rounds the personal experience obtained with the given co-player becomes more decisive⁶³.

The widespread tendency to judge actions between third parties, and the readiness for cooperation combined with altruistic punishment (also known as strong reciprocity⁷⁰) has been neatly captured in an experiment involving three players^{71,72}. First, players A and B engage in one round of a Prisoner's Dilemma game. Then player C has the possibility to mete out costly punishment on A and B. Defectors are often punished, although this reduces the endowment of the punisher. It would be interesting to see whether observers of a repeated Prisoner's Dilemma game are using a similar level of punishment, or whether they reduce it because wronged players have the opportunity of avenging themselves.

Indirect reciprocity and public goods games are also closely connected. For example, donors are more generous if they learn that the recipient has recently made a donation to a charitable institution⁷³. An even more remarkable effect was found in an experiment alternating rounds of the Public Goods Game with rounds of the indirect reciprocity game⁷⁴. It is known that many players show an initial willingness to contribute to the public good a substantial amount of their endowment, but this willingness often vanishes within a few rounds. This is not the case if indirect reciprocity games are sandwiched between the rounds of the Public Goods Game. If players are informed about their recipient's action in the Public Goods Game, they tend to be more generous towards recipients who contributed much. Conversely, players are more willing to contribute to the public good if they know that this will be communicated before the start of the indirect reciprocity game. The contributions to the public good do not deteriorate from one round to the next. The donations in the indirect reciprocity game, which are channelled towards those who contributed much to the public good, can be viewed as rewards.

Whereas most experiments, in indirect reciprocity, were motivated by models, this last experiment led to a model^{75,76}. A discriminating strategy, which defects in all rounds of the indirect reciprocity game if the recipient is known to have defected in the Public Goods Game, can guarantee stable cooperation. Because this discriminating strategy distinguishes between justified and unjustified defection, it is effectively a non-costly form of punishing free riders in the Public Goods Game.

Future directions

Thus indirect reciprocity based on reputation serves as a link between diverse forms of cooperative interaction. The moralistic assessment of the other members in the population, even if they are observed only at a distance, provides a powerful tool for channelling support towards those who collaborate, and an incentive to join group efforts.

Box 5 | Social viscosity

Altruism towards genetic relatives can evolve by kin selection provided that Hamilton's rule¹⁴ holds: the coefficient of relatedness, R , between the donor and the recipient has to exceed the cost-to-benefit ratio of the altruistic act:

$$R > c/b.$$

As Haldane has said, 'I will jump into the river to save two brothers or eight cousins.' The probability that brothers share a 'selfish gene' is 1/2; the same probability for cousins is 1/8. Kin selection works in 'viscous' populations in which chances are high that neighbours are genetic relatives.

For indirect reciprocity a similar rule holds^{36,86}: the probability, q , of knowing the social score of another person must exceed the cost-to-benefit ratio:

$$q > c/b.$$

The role of genetic relatedness that is crucial for kin selection is replaced by social acquaintanceship. In a fluid population, in which most interactions are anonymous and people have no possibility of monitoring the social score of others, indirect reciprocity has no chance. In a socially viscous population, in which people know each other's reputation, cooperation by indirect reciprocity can thrive.

The exploration of the links between indirect reciprocity and other game theoretical models for cooperative interaction promises to offer further opportunities for a better understanding of human traits (Boxes 4 and 5). Future theoretical and experimental work on indirect reciprocity is likely to go beyond the context of economic interactions in the narrow sense and to address such issues as the physiological correlate of trust and decision-making, the emergence of language capabilities and moral norms, subliminal effects framing our beliefs, and the pervasive role of individual reputations and social prejudice.

1. Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
2. Alexander, R. D. *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).
3. Binmore, K. G. *Game Theory and the Social Contract* (MIT Press, Cambridge, Massachusetts, 1994).
4. Seabright, P. *The Company of Strangers: A Natural History of Economic Life* (Princeton Univ. Press, Princeton, 2004).
5. Bolton, G. E., Katok, E. & Ockenfels, A. How effective are electronic reputation mechanisms? An experimental investigation. *Manage. Sci.* **50**, 1587–1602 (2004).
6. Keser, C. Trust and reputation building in e-commerce. IBM Watson Research Center, CIRANO working paper no. 2002s-75 (http://www.cirano.qc.ca/en/publication_detail.php?id=2002s-75) (2002).
7. Dellarocas, C. Sanctioning reputation mechanisms in online trading environments with moral hazard. MIT Sloan School of Management working paper no. 4297-03 (<http://ssrn.com/abstract=393043>) (2003).
8. Bolton, G., Katok, E. & Ockenfels, A. in *Applications of Supply Chain Management and e-Commerce Research* (eds Geunes, J., Akcali, E., Pardalos, P. M., Romeijn, H. E. & Shen, Z.-J.) 195–216 (Springer, Dordrecht, 2005).
9. Resnick, P., Zeckhauser, R., Friedman, E. & Kuwabara, K. Reputation systems. *Commun. ACM* **43**, 45–48 (2000).
10. Lucking-Reiley, D., Bryan, D., Prasad, N. & Reeves, D. *Pennies from eBay: The Determinants of Price in Online Auctions*. Vanderbilt Univ., Univ. Arizona working paper (1999).
11. Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford, 1997).
12. Wilson, E. O. *Sociobiology* (Harvard Univ. Press, Cambridge, Massachusetts, 1975).
13. Trivers, R. *Social Evolution* (Benjamin Cummings, Menlo Park, 1985).
14. Hamilton, W. D. *Narrow Roads of Gene Land* Vol. 1 (Freeman, New York, 1996).
15. Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758 (2003).
16. Rilling, J. K. et al. A neural basis for social cooperation. *Neuron* **35**, 395–405 (2002).
17. Fehr, E. The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).
18. Dunbar, R. *Grooming, Gossip and the Evolution of Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1996).
19. Brown, D. E. *Human Universals* (McGraw-Hill, New York, 1991).
20. Whiten, A. & Byrne, R. W. (eds) *Machiavellian Intelligence II: Extensions and Evaluations* (Cambridge Univ. Press, Cambridge, UK, 1997).
21. Axelrod, R. *The Evolution of Cooperation* (Reprinted by Penguin, Harmondsworth, 1989.) (Basic Books, New York, 1984).
22. Nowak, M. A. & Sigmund, K. The alternating prisoner's dilemma. *J. Theor. Biol.* **168**, 219–226 (1994).
23. Frean, M. R. The prisoner's dilemma without synchrony. *Proc. R. Soc. Lond. B* **257**, 75–79 (1994).
24. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
25. Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **50**, 533–554 (1986).
26. Binmore, K. *Fun and Games: A Text on Game Theory* (Heath, Lexington, Massachusetts, 1992).
27. Rosenthal, R. W. Sequences of games with varying opponents. *Econometrica* **47**, 1353–1366 (1979).
28. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59**, 63–80 (1992).
29. Ellison, G. Cooperation in the prisoner's dilemma with anonymous random matching. *Rev. Econ. Stud.* **61**, 567–588 (1994).
30. Okuno-Fujiwara, M. & Postlewaite, A. Social norms in matching games. *Games Econ. Behav.* **9**, 79–109 (1995).
31. Nowak, M. A. & Sigmund, K. Evolutionary dynamics of biological games. *Science* **303**, 793–799 (2004).
32. Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
33. Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
34. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
35. Nowak, M. A. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58 (1993).
36. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
37. Wedekind, C. Give and ye shall be recognized. *Science* **280**, 2070–2071 (1998).
38. Ferrière, R. Help and you shall be helped. *Nature* **393**, 517–519 (1998).
39. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocation. *Proc. R. Soc. Lond. B* **268**, 745–753 (2001).
40. Panchanathan, K. & Boyd, R. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
41. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574 (1998).
42. Ohtsuki, H. & Iwasa, Y. How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
43. Dawes, R. M. Social dilemmas. *Annu. Rev. Psychol.* **31**, 169–193 (1980).
44. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
45. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
46. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
47. Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10757–10762 (2001).
48. Sugden, R. *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, 1986).
49. Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* (in the press).
50. Lotem, A., Fishman, M. A. & Stone, L. Evolution of cooperation between individuals. *Nature* **400**, 226–227 (1999).
51. Lotem, A., Fishman, M. A. & Stone, L. Evolution of unconditional altruism through signalling benefits. *Proc. R. Soc. Lond. B* **270**, 199–205 (2002).
52. Fishman, M. A., Lotem, A. & Stone, L. Heterogeneity stabilises reciprocal altruism interaction. *J. Theor. Biol.* **209**, 87–95 (2001).
53. Fishman, M. A. Indirect reciprocity among imperfect individuals. *J. Theor. Biol.* **225**, 285–292 (2003).
54. Brandt, H. & Sigmund, K. The logic of reprobation: Assessment and action rules for indirect reciprocity. *J. Theor. Biol.* **231**, 475–486 (2004).
55. Mohtashemi, M. & Mui, L. Evolution of indirect reciprocity by social information: The role of trust and reputation in evolution of altruism. *J. Theor. Biol.* **223**, 523–531 (2003).
56. Brandt, H. & Sigmund, K. Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670 (2005).
57. Takahashi, N. & Mashima, R. The emergence of indirect reciprocity: Is the standing strategy the answer? Hokkaido Univ. working paper no. 29 (<http://lynx.let.hokudai.ac.jp/COE21/pdf/029.zip>) (2003).
58. Wilson, D. S. A theory of group selection. *Proc. Natl Acad. Sci. USA* **72**, 143–146 (1975).
59. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
60. Seinen, I. & Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.* (in the press).
61. Haley, K. J. & Kessler, D. M. T. Nobody is watching?: Subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26**, 245–256 (2005).
62. Bolton, G. E., Katok, E. & Ockenfels, A. Cooperation among strangers with limited information about reputation. *J. Public Econ.* **89**, 1457–1468 (2005).
63. Wedekind, C. & Braithwaite, V. A. The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12**, 1012–1015 (2002).
64. Engelmann, D. & Fischbacher, U. Indirect reciprocity and strategic reputation building in an experimental helping game. Univ. Zürich working paper no. 132 (<http://www.iew.unizh.ch/wp/iewwp132.pdf>) (2002).
65. Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268**, 2495–2501 (2001).
66. Greiner, B. & Levati, M. V. Indirect reciprocity in cyclical networks: An experimental study. Max Planck Institute, Jena, working paper no. 2003-15 (<ftp://papers.mpiw-jena.mpg.de/esi/discussionpapers/2003-15.pdf>) (2003).
67. Dufwenberg, M., Gneezy, U., Güth, W. & van Damme, E. Direct vs indirect reciprocity: An experiment. *Homo Oecon.* **18**, 19–30 (2001).
68. Güth, W., Königstein, M., Marchand, N. & Nehring, K. Trust and reciprocity in the investment game with indirect reward. *Homo Oecon.* **18**, 241–262 (2001).
69. Pollock, G. B. & Dugatkin, L. A. Reciprocity and the evolution of reputation. *J. Theor. Biol.* **159**, 25–37 (1992).
70. Gintis, H. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction* (Princeton Univ. Press, Princeton, 2000).
71. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190 (2004).

72. Fehr, E. & Fischbacher, U. Third party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
73. Milinski, M., Semmann, D. & Krambeck, H. J. Donors in charity gain in both indirect reciprocity and political reputation. *Proc. R. Soc. Lond. B* **269**, 881–883 (2002).
74. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
75. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free-rider problem. *Nature* **432**, 499–502 (2004).
76. Fehr, E. Don't lose your reputation. *Nature* **432**, 449–450 (2004).
77. Boyd, R. & Richerson, P. J. The evolution of indirect reciprocity. *Soc. Networks* **11**, 213–236 (1989).
78. Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M. & Bonhoeffer, S. Evolution of cooperation by generalized reciprocity. *Proc. R. Soc. B* **272**, 1115–1120 (2005).
79. Brandt, H. & Sigmund, K. The good, the bad and the discriminator. *J. Theor. Biol.* (in the press).
80. Imhof, L., Fudenberg, D. & Nowak, M. A. Evolutionary cycles of cooperation and defection. *Proc. Natl Acad. Sci. USA* **102**, 10797–10800 (2005).
81. Camerer, C. E. *Behavioral Game Theory* (Princeton Univ. Press, Princeton, 2003).
82. Colman, A. M. *Game Theory and its Applications in the Social and Biological Sciences* (Butterworth-Heinemann, Oxford, 1995).
83. Arrow, K. *The Limits of Organization* (Norton, New York, 1974).
84. Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U. & Fehr, E. Oxytocin increases trust in humans. *Nature* **435**, 673–676 (2005).
85. Milgrom, P., North, D. & Weingast, B. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Econ. Politics* **2**, 1–23 (1990).
86. Suzuki, Y. & Toquenaga, Y. Effects of information and group structure on evolution of altruism: Analysis of two-score model by covariance and contextual analyses. *J. Theor. Biol.* **232**, 191–203 (2005).
87. Fudenberg, D. & Maskin, E. Evolution of cooperation in noisy repeated games. *American Economic Review* **80**, 274–279 (1990).

Acknowledgements Support from the John Templeton Foundation is gratefully acknowledged. The Program for Evolutionary Dynamics at Harvard University is sponsored by Jeffrey Epstein.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to K.S. (karl.sigmund@univie.ac.at).

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Despite the ever-accelerating pace of biomedical research, the root causes of common human diseases remain largely unknown, preventative measures are generally inadequate, and available treatments are seldom curative. Family history is one of the strongest risk factors for nearly all diseases—including cardiovascular disease, cancer, diabetes, autoimmunity, psychiatric illnesses and many others—providing the tantalizing but elusive clue that inherited genetic variation has an important role in the pathogenesis of disease. Identifying the causal genes and variants would represent an important step in the path towards improved prevention, diagnosis and treatment of disease.

More than a thousand genes for rare, highly heritable ‘mendelian’ disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease. Common disorders, in contrast, have proven much more challenging to study, as they are thought to be due to the combined effect of many different susceptibility DNA variants interacting with environmental factors.

Studies of common diseases have fallen into two broad categories: family-based linkage studies across the entire genome, and population-based association studies of individual candidate genes. Although there have been notable successes, progress has been slow due to the inherent limitations of the methods; linkage analysis has low power except when a single locus explains a substantial fraction of disease, and association studies of one or a few candidate genes examine only a small fraction of the ‘universe’ of sequence variation in each patient.

A comprehensive search for genetic influences on disease would involve examining all genetic differences in a large number of affected individuals and controls. It may eventually become possible to accomplish this by complete genome resequencing. In the meantime, it is increasingly practical to systematically test common genetic variants for their role in disease; such variants explain much of the genetic diversity in our species, a consequence of the historically small size and shared ancestry of the human population.

Recent experience bears out the hypothesis that common variants have an important role in disease, with a partial list of validated examples including *HLA* (autoimmunity and infection)¹, *APOE4* (Alzheimer’s disease, lipids)², Factor V^{Leiden} (deep vein thrombosis)³, *PPARG* (encoding PPAR γ ; type 2 diabetes)^{4,5}, *KCNJ11* (type 2

diabetes)⁶, *PTPN22* (rheumatoid arthritis and type 1 diabetes)^{7,8}, insulin (type 1 diabetes)⁹, *CTLA4* (autoimmune thyroid disease, type 1 diabetes)¹⁰, *NOD2* (inflammatory bowel disease)^{11,12}, complement factor H (age-related macular degeneration)^{13–15} and *RET* (Hirschsprung disease)^{16,17}, among many others.

Systematic studies of common genetic variants are facilitated by the fact that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as linkage disequilibrium (LD); a particular combination of alleles along a chromosome is termed a haplotype.

LD exists because of the shared ancestry of contemporary chromosomes. When a new causal variant arises through mutation—whether a single nucleotide change, insertion/deletion, or structural alteration—it is initially tethered to a unique chromosome on which it occurred, marked by a distinct combination of genetic variants. Recombination and mutation subsequently act to erode this association, but do so slowly (each occurring at an average rate of about 10^{-8} per base pair (bp) per generation) as compared to the number of generations (typically 10^4 to 10^5) since the mutational event.

The correlations between causal mutations and the haplotypes on which they arose have long served as a tool for human genetic research: first finding association to a haplotype, and then subsequently identifying the causal mutation(s) that it carries. This was pioneered in studies of the *HLA* region, extended to identify causal genes for mendelian diseases (for example, cystic fibrosis¹⁸ and diastrophic dysplasia¹⁹), and most recently for complex disorders such as age-related macular degeneration^{13–15}.

Early information documented the existence of LD in the human genome^{20,21}; however, these studies were limited (for technical reasons) to a small number of regions with incomplete data, and general patterns were challenging to discern. With the sequencing of the human genome and development of high-throughput genomic methods, it became clear that the human genome generally displays more LD²² than under simple population genetic models²³, and that LD is more varied across regions, and more segmentally structured^{24–30}, than had previously been supposed. These observations indicated that LD-based methods would generally have great value (because nearby SNPs were typically correlated with many of their neighbours), and also that LD relationships would

*Lists of participants and affiliations appear at the end of the paper.

Centre	Chromosomes	Technology
RIKEN	5, 11, 14, 15, 16, 17, 19	Third Wave Invader
Wellcome Trust Sanger Institute	1, 6, 10, 13, 20	Illumina BeadArray
McGill University and Génome Québec Innovation Centre	2, 4p	Illumina BeadArray
Chinese HapMap Consortium*	3, 8p, 21	Sequenom MassExtend, Illumina BeadArray
Illumina	8q, 9, 18q, 22, X	Illumina BeadArray
Broad Institute of Harvard and MIT	4q, 7q, 18p, Y, mtDNA	Sequenom MassExtend, Illumina BeadArray
Baylor College of Medicine with ParAllele BioScience	12	ParAllele MIP
University of California, San Francisco, with Washington University in St Louis	7p	PerkinElmer AcycloPrime-FP
Perlegen Sciences	5 Mb (ENCODE) on 2, 4, 7, 8, 9, 12, 18 in CEU	High-density oligonucleotide array

also Supplementary Table 1). All genotyping centres produced high-quality data (accuracy more than 99% in the blind QA exercises, Supplementary Tables 2 and 3), and missing data were not biased against heterozygotes. The Supplementary Information contains the full details of these efforts.

Although SNP selection was generally agnostic to functional annotation, 11,500 non-synonymous cSNPs (SNPs in coding regions of genes where the different SNP alleles code for different amino acids in the protein) were successfully typed in Phase I. (An effort was made to prioritize cSNPs in Phase I in choosing SNPs for each 5-kb region; all known non-synonymous cSNPs were attempted as part of Phase II.)

Across the ten ENCODE regions (Table 2), the density of SNPs was approximately tenfold higher as compared to the genome-wide map: 17,944 SNPs across the 5 megabases (Mb) (one per 279 bp).

More than 1.3 million SNP genotyping assays were attempted (Table 3) to generate the Phase I data on more than 1 million SNPs. The 0.3 million SNPs not part of the Phase I data set include 73,652 that passed QC filters but were monomorphic in all 269 samples. The remaining SNPs failed the QC filters in one or more analysis panels mostly because of inadequate completeness, non-mendelian inheritance, deviations from Hardy–Weinberg equilibrium, discrepant genotypes among duplicates, and data transmission discrepancies.

SNPs on the Phase I map are evenly spaced, except on Y and mtDNA. The Phase I data include a successful, common SNP every 5 kb across most of the genome in each analysis panel

(Supplementary Fig. 1): only 3.3% of inter-SNP distances are longer than 10 kb, spanning 11.9% of the genome (Fig. 2; see also Supplementary Fig. 2). One exception is the X chromosome (Supplementary Fig. 1), where a much higher proportion of attempted SNPs were rare or monomorphic, and thus the density of common SNPs is lower.

Two intentional exceptions to the regular spacing of SNPs on the physical map were the mitochondrial chromosome (mtDNA), which does not undergo recombination, and the non-recombining portion of chromosome Y. On the basis of the 168 successful, polymorphic SNPs, each HapMap sample fell into one of 15 (of the 18 known) mtDNA haplogroups³⁴ (Table 4). A total of 84 SNPs that characterize the unique branches of the reference Y genealogical tree^{35–37} were genotyped on the HapMap samples. These SNPs assigned each Y chromosome to 8 (of the 18 major) Y haplogroups previously described (Table 4).

Highly accurate phasing of long-range chromosomal haplotypes. Despite having collected data in diploid individuals, the inclusion of parent–offspring trios and the use of computational methods made it possible to determine long-range phased haplotypes of extremely high quality for each individual. These computational algorithms take advantage of the observation that because of LD, relatively few of the large number of possible haplotypes consistent with the genotype data actually occur in population samples.

The project compared a variety of algorithms for phasing haplotypes from unrelated individuals and trios³⁸, and applied the algorithm that proved most accurate (an updated version of PHASE³⁹)

Table 2 | ENCODE project regions and genotyping

Region name	Chromosome band	Genomic interval (NCBI) (base numbers)†	Gene density (%)‡	Conservation score (%)§	Pedigree-based recombination rate (cM Mb ⁻¹)	Population-based recombination rate (cM Mb ⁻¹)¶	G+C content#	Available SNPs			Successfully genotyped SNPs††	Sequencing centre/ genotyping centre(s)‡‡
								dbSNP☆	Sequence**	Total		
ENr112	2p16.3	51,633,239–52,133,238	0	3.8	0.8	0.9	0.35	1,570	1,762	3,332	2,275	Broad/McGill-GQIC
ENr131	2q37.1	234,778,639–235,278,638	4.6	1.3	2.2	2.5	0.43	1,736	1,259	2,995	1,910	Broad/McGill-GQIC
ENr113	4q26	118,705,475–119,205,474	0	3.9	0.6	0.9	0.35	1,444	2,053	3,497	2,201	Broad/Broad
ENm010	7p15.2	26,699,793–27,199,792	5.0	22.0	0.9	0.9	0.44	1,220	1,795	3,015	1,271	Baylor/UCSF-WU,
ENm013*	7q21.13	89,395,718–89,895,717	5.5	4.4	0.4	0.5	0.38	1,394	1,917	3,311	1,807	Broad/Broad
ENm014*	7q31.33	126,135,436–126,632,577	2.9	11.2	0.4	0.9	0.39	1,320	1,664	2,984	1,966	Broad/Broad
ENr321	8q24.11	118,769,628–119,269,627	3.2	11.4	0.6	1.1	0.41	1,430	1,508	2,938	1,758	Baylor/Illumina
ENr232	9q34.11	127,061,347–127,561,346	5.9	8.3	2.7	2.6	0.52	1,444	1,523	2,967	1,324	Baylor/Illumina
ENr123	12q12	38,626,477–39,126,476	3.1	1.7	0.3	0.8	0.36	1,877	1,379	3,256	1,792	Baylor /
ENr213	18q12.1	23,717,221–24,217,220	0.9	7.4	1.2	0.9	0.37	1,330	1,459	2,789	1,640	Baylor/Illumina
Total	–	–	–	–	–	–	–	14,765	16,319	31,084	17,944	–

McGill-GQIC, McGill University and Génome Québec Innovation Centre.

* These regions were truncated to 500 kb for resequencing.

† Sequence build 34 coordinates.

‡ Gene density is defined as the percentage of bases covered either by Ensembl genes or human mRNA best BLAT alignments in the UCSC Genome Browser database.

§ Non-exonic conservation with mouse sequence was measured by taking 125 base non-overlapping sub-windows inside the 500,000 base windows. Sub-windows with less than 75% of their bases in a mouse alignment were discarded. Of the remaining sub-windows, those with at least 80% base identity were used to calculate the conservation score. The mouse alignments in regions corresponding to the following were discarded: Ensembl genes, all GenBank mRNA Blastz alignments, FGENESH++ gene predictions, Twinscan gene predictions, spliced EST alignments, and repeats.

|| The pedigree-based sex-averaged recombination map is from deCODE Genetics⁴⁸.

¶ Recombination rate based on estimates from LDhat⁴⁶.

G + C content calculated from the sequence of the stated coordinates from sequence build 34.

☆ SNPs in dbSNP build 121 at the time the ENCODE resequencing began and SNPs added to dbSNP in builds 122–125 independent of the resequencing.

** New SNPs discovered through the resequencing reported here (not found by other means in builds 122–125).

†† SNPs successfully genotyped in all analysis panels (YRI, CEU, CHB + JPT).

‡‡ Perlegen genotyped a subset of SNPs in the CEU samples.

Table 3 | HapMap Phase I genotyping success measures

SNP categories	Analysis panel		
	YRI	CEU	CHB + JPT
Assays submitted	1,273,716	1,302,849	1,273,703
Passed QC filters	1,123,296 (88%)	1,157,650 (89%)	1,134,726 (89%)
Did not pass QC filters*	150,420 (12%)	145,199 (11%)	138,977 (11%)
> 20% missing data	98,116 (65%)	107,626 (74%)	93,710 (67%)
> 1 duplicate inconsistent	7,575 (5%)	6,254 (4%)	10,725 (8%)
> 1 mendelian error	22,815 (15%)	13,600 (9%)	0 (0%)
< 0.001 Hardy–Weinberg <i>P</i> -value	12,052 (8%)	9,721 (7%)	16,176 (12%)
Other failures†	23,478 (16%)	17,692 (12%)	23,722 (17%)
Non-redundant (unique) SNPs	1,076,392	1,104,980	1,087,305
Monomorphic	156,290 (15%)	234,482 (21%)	268,325 (25%)
Polymorphic	920,102 (85%)	870,498 (79%)	818,980 (75%)
All analysis panels			
Unique QC-passed SNPs	1,156,772		
Passed in one analysis panel	52,204 (5%)		
Passed in two analysis panels	97,231 (8%)		
Passed in three analysis panels	1,007,337 (87%)		
Monomorphic across three analysis panels	75,997		
Polymorphic in all three analysis panels	682,397		
MAF ≥ 0.05 in at least one of three analysis panels	877,351		

* Out of 95 samples in CEU, YRI; 94 samples in CHB + JPT.
† ‘Other failures’ includes SNPs with discrepancies during the data transmission process. Some SNPs failed in more than one way, so these percentages add up to more than 100%.

separately to each analysis panel. (Phased haplotypes are available for download at the Project website.) We estimate that ‘switch’ errors—where a segment of the maternal haplotype is incorrectly joined to the paternal—occur extraordinarily rarely in the trio samples (every 8 Mb in CEU; 3.6 Mb in YRI). The switch rate is higher in the CHB+JPT samples (one per 0.34 Mb) due to the lack of information from parent–offspring trios, but even for the unrelated individuals, statistical reconstruction of haplotypes is remarkably accurate.

Estimating properties of SNP discovery and dbSNP. Extensive sequencing and genotyping in the ENCODE regions characterized

the false-positive and false-negative rates for dbSNP, as well as polymerase chain reaction (PCR)-based resequencing (see Methods). These data reveal two important conclusions: first, that PCR-based sequencing of diploid samples may be biased against very rare variants (that is, those seen only as a single heterozygote), and second, that the vast majority of common variants are either represented in dbSNP, or show tight correlation to other SNPs that are in dbSNP (Fig. 3).

Allele frequency distributions within population samples. The underlying allele frequency distributions for these samples are best

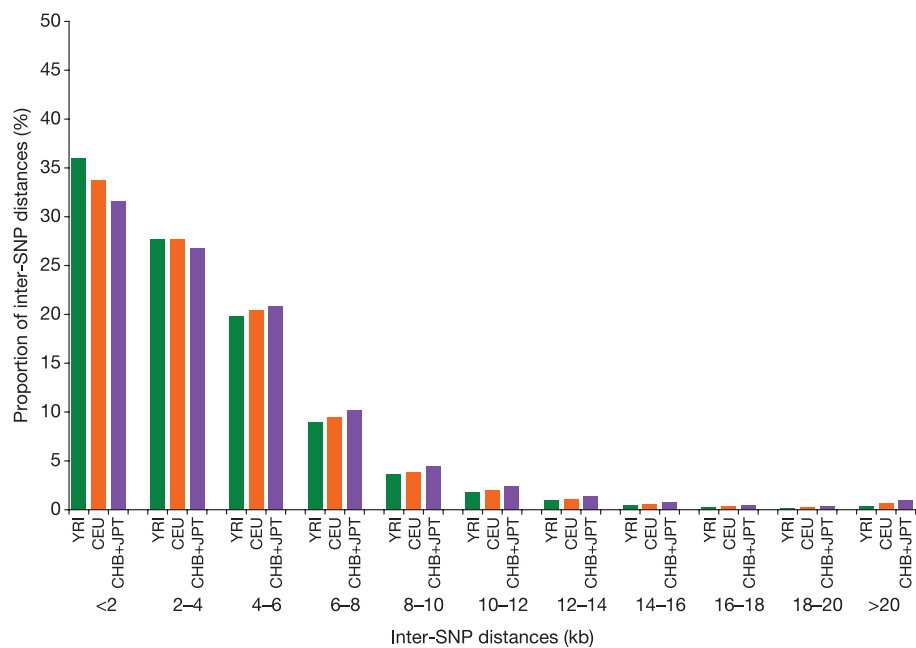


Figure 2 | Distribution of inter-SNP distances. The distributions are shown for each analysis panel for the HapMappable genome (defined in the Methods), for all common SNPs (with MAF ≥ 0.05).

estimated from the ENCODE data, where deep sequencing reduces bias due to SNP ascertainment. Consistent with previous studies, most SNPs observed in the ENCODE regions are rare: 46% had $MAF < 0.05$, and 9% were seen in only a single individual (Fig. 4). Although most varying sites in the population are rare, most heterozygous sites within any individual are due to common SNPs. Specifically, in the ENCODE data, 90% of heterozygous sites in each individual were due to common variants (Fig. 4). With ever-deeper sequencing of DNA samples the number of rare variants will rise linearly, but the vast majority of heterozygous sites in each person will be explained by a limited set of common SNPs now contained (or captured through LD) in existing databases (Fig. 3).

Consistent with previous descriptions, the CEU, CHB and JPT samples show fewer low frequency alleles when compared to the YRI samples (Fig. 5), a pattern thought to be due to bottlenecks in the history of the non-YRI populations.

In contrast to the ENCODE data, the distribution of allele frequencies for the genome-wide data is flat (Fig. 5), with much more similarity in the distributions observed in the three analysis panels. These patterns are well explained by the inherent and intentional bias in the rules used for SNP selection: we prioritized using validated SNPs in order to focus resources on common (rather than rare or false positive) candidate SNPs from the public databases. For a fuller discussion of ascertainment issues, including a shift in frequencies over time and an excess of high-frequency derived alleles due to inclusion of chimpanzee data in determination of double-hit status, see the Supplementary Information (Supplementary Fig. 3). **SNP allele frequencies across population samples.** Of the 1.007 million SNPs successfully genotyped and polymorphic across the three analysis panels, only a subset were polymorphic in any given panel: 85% in YRI, 79% in CEU, and 75% in CHB+JPT. The joint distribution of frequencies across populations is presented in Fig. 6 (for the ENCODE data) and Supplementary Fig. 4 (for the genome-wide map). We note the similarity of allele frequencies in the CHB and JPT samples, which motivates analysing them jointly as a single analysis panel in the remainder of this report.

Table 4 | mtDNA and Y chromosome haplogroups

MtDNA haplogroup	DNA sample*			
	YRI (60)	CEU (60)	CHB (45)	JPT (44)
L1	0.22	-	-	-
L2	0.35	-	-	-
L3	0.43	-	-	-
A	-	-	0.13	0.04
B	-	-	0.33	0.30
C	-	-	0.09	0.07
D	-	-	0.22	0.34
M/E	-	-	0.22	0.25
H	-	0.45	-	-
V	-	0.07	-	-
J	-	0.08	-	-
T	-	0.12	-	-
K	-	0.03	-	-
U	-	0.23	-	-
W	-	0.02	-	-

Y chromosome haplogroup	DNA sample*			
	YRI (30)	CEU (30)	CHB (22)	JPT (22)
E1	0.07	-	-	-
E3a	0.93	-	-	-
F, H, K	-	0.03	0.23	0.14
I	-	0.27	-	-
R1	-	0.70	-	-
C	-	-	0.09	0.09
D	-	-	-	0.45
NO	-	-	0.68	0.32

*Number of chromosomes sampled is given in parentheses.

A simple measure of population differentiation is Wright's F_{ST} , which measures the fraction of total genetic variation due to between-population differences⁴⁰. Across the autosomes, F_{ST} estimated from the full set of Phase I data is 0.12, with CEU and CHB+JPT showing the lowest level of differentiation ($F_{ST} = 0.07$), and YRI and CHB+JPT the highest ($F_{ST} = 0.12$). These values are slightly higher than previous reports⁴¹, but differences in the types of variants (SNPs versus microsatellites) and the samples studied make comparisons difficult.

As expected, we observed very few fixed differences (that is, cases in which alternate alleles are seen exclusively in different analysis panels). Across the 1 million SNPs genotyped, only 11 have fixed differences between CEU and YRI, 21 between CEU and CHB+JPT, and 5 between YRI and CHB+JPT, for the autosomes.

The extent of differentiation is similar across the autosomes, but higher on the X chromosome ($F_{ST} = 0.21$). Interestingly, 123 SNPs on the X chromosome were completely differentiated between YRI and CHB+JPT, but only two between CEU and YRI and one between CEU and CHB+JPT. This seems to be largely due to a single region near the centromere, possibly indicating a history of natural

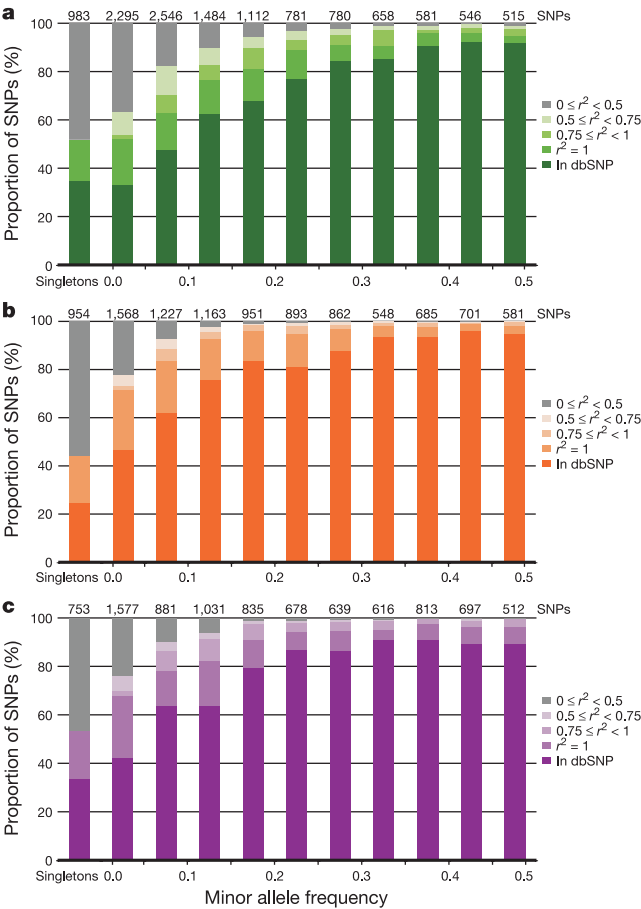


Figure 3 | Allele frequency and completeness of dbSNP for the ENCODE regions. **a–c**, The fraction of SNPs in dbSNP, or with a proxy in dbSNP, are shown as a function of minor allele frequency for each analysis panel (**a**, YRI; **b**, CEU; **c**, CHB+JPT). Singletons refer to heterozygotes observed in a single individual, and are broken out from other SNPs with $MAF < 0.05$. Because all ENCODE SNPs have been deposited in dbSNP, for this figure we define a SNP as ‘in dbSNP’ if it would be in dbSNP build 125 independent of the HapMap ENCODE resequencing project. All remaining SNPs (not in dbSNP) were discovered only by ENCODE resequencing; they are categorized by their correlation (r^2) to those in dbSNP. Note that the number of SNPs in each frequency bin differs among analysis panels, because not all SNPs are polymorphic in all analysis panels.

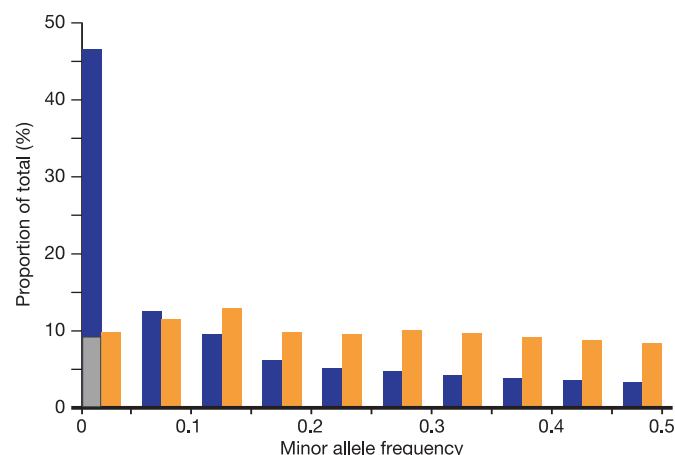


Figure 4 | Minor allele frequency distribution of SNPs in the ENCODE data, and their contribution to heterozygosity. This figure shows the polymorphic SNPs from the HapMap ENCODE regions according to minor allele frequency (blue), with the lowest minor allele frequency bin (<0.05) separated into singletons (SNPs heterozygous in one individual only, shown in grey) and SNPs with more than one heterozygous individual. For this analysis, MAF is averaged across the analysis panels. The sum of the contribution of each MAF bin to the overall heterozygosity of the ENCODE regions is also shown (orange).

selection at this locus (see below; M. L. Freedman *et al.*, personal communication).

Haplotype sharing across populations. We next examined the extent to which haplotypes are shared across populations. We used a hidden Markov model in which each haplotype is modelled in turn as an imperfect mosaic of other haplotypes (see Supplementary Information)⁴². In essence, the method infers probabilistically which other haplotype in the sample is the closest relative (nearest neighbour) at each position along the chromosome.

Unsurprisingly, the nearest neighbour most often is from the same analysis panel, but about 10% of haplotypes were found most closely to match a haplotype in another panel (Supplementary Fig. 5). All individuals have at least some segments over which the nearest neighbour is in a different analysis panel. These results indicate that although analysis panels are characterized both by different haplotype frequencies and, to some extent, different combinations of alleles, both common and rare haplotypes are often shared across populations.

Properties of LD in the human genome

Traditionally, descriptions of LD have focused on measures calculated between pairs of SNPs, averaged as a function of physical distance. Examples of such analyses for the HapMap data are presented in Supplementary Fig. 6. After adjusting for known confounders such as sample size, allele frequency distribution, marker density, and length of sampled regions, these data are highly similar to previously published surveys⁴³.

Because LD varies markedly on scales of 1–100 kb, and is often discontinuous rather than declining smoothly with distance, averages obscure important aspects of LD structure. A fuller exploration of the fine-scale structure of LD offers both insight into the causes of LD and understanding of its application to disease research. **LD patterns are simple in the absence of recombination.** The most natural path to understanding LD structure is first to consider the simplest case in which there is no recombination (or gene conversion), and then to add recombination to the model. (For simplicity we ignore genotyping error and recurrent mutation in this discussion, both of which seem to be rare in these data.)

In the absence of recombination, diversity arises solely through mutation. Because each SNP arose on a particular branch of the genealogical tree relating the chromosomes in the current populations, multiple haplotypes are observed. SNPs that arose on the same branch of the genealogy are perfectly correlated in the sample, whereas SNPs that occurred on different branches have imperfect correlations, or no correlation at all.

We illustrate these concepts using empirical genotype data from 36 adjacent SNPs in an ENCODE region (ENr131.2q37), selected because no obligate recombination events were detectable among them in CEU (Fig. 7). (We note that the lack of obligate recombination events in a small sample does not guarantee that no recombinants have occurred, but it provides a good approximation for illustration.)

In principle, 36 such SNPs could give rise to 2^{36} different haplotypes. Even with no recombination, gene conversion or recurrent mutation, up to 37 different haplotypes could be formed. Despite this great potential diversity, only seven haplotypes are observed (five seen more than once) among the 120 parental CEU chromosomes studied, reflecting shared ancestry since their most recent common ancestor among apparently unrelated individuals.

In such a setting, it is easy to interpret the two most common pairwise measures of LD: D' and r^2 . (See the Supplementary Information for fuller definitions of these measures.) D' is defined to be 1 in the absence of obligate recombination, declining only due to recombination or recurrent mutation²⁷. In contrast, r^2 is simply

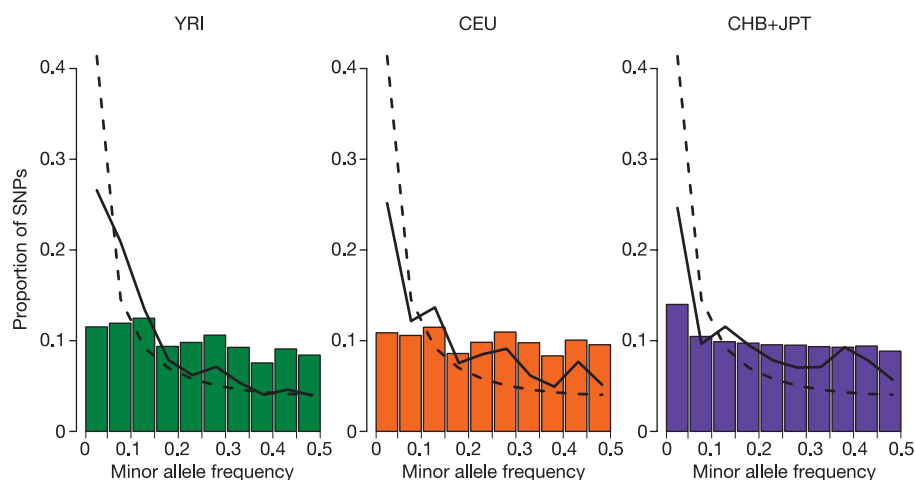


Figure 5 | Allele frequency distributions for autosomal SNPs. For each analysis panel we plotted (bars) the MAF distribution of all the Phase I SNPs with a frequency greater than zero. The solid line shows the MAF

distribution for the ENCODE SNPs, and the dashed line shows the MAF distribution expected for the standard neutral population model with constant population size and random mating without ascertainment bias.

the squared correlation coefficient between the two SNPs. Thus, r^2 is 1 when two SNPs arose on the same branch of the genealogy and remain undisrupted by recombination, but has a value less than 1 when SNPs arose on different branches, or if an initially strong correlation has been disrupted by crossing over.

In this region, $D' = 1$ for all marker pairs, as there is no evidence of historical recombination. In contrast, and despite great simplicity of haplotype structure, r^2 values display a complex pattern, varying from 0.0003 to 1.0, with no relationship to physical distance. This makes sense, however, because without recombination, correlations among SNPs depend on the historical order in which they arose, not the physical order of SNPs on the chromosome.

Most importantly, the seeming complexity of r^2 values can be deconvolved in a simple manner: only seven different SNP configurations exist in this region, with all but two chromosomes matching five common haplotypes, which can be distinguished from each other by typing a specific set of four SNPs. That is, only a small minority of sites need be examined to capture fully the information in this region.

Variation in local recombination rates is a major determinant of LD. Recombination in the ancestors of the current population has typically disrupted the simple picture presented above. In the human genome, as in yeast⁴⁴, mouse⁴⁵ and other genomes, recombination rates typically vary dramatically on a fine scale, with hotspots of recombination explaining much crossing over in each region²⁸. The generality of this model has recently been demonstrated through computational methods that allow estimation of recombination rates (including hotspots and coldspots) from genotype data^{46,47}.

The availability of nearly complete information about common DNA variation in the ENCODE regions allowed a more precise estimation of recombination rates across large regions than in any previous study. We estimated recombination rates and identified recombination hotspots in the ENCODE data, using methods previously described⁴⁶ (see Supplementary Information for details). Hotspots are short regions (typically spanning about 2 kb) over which recombination rates rise dramatically over local background rates.

Whereas the average recombination rate over 500 kb across the human genome is about 0.5 cM⁴⁸, the estimated recombination rate across the 500-kb ENCODE regions varied nearly tenfold, from a minimum of 0.19 cM (ENm013.7q21.13) to a maximum of 1.25 cM (ENr232.9q34.11). Even this tenfold variation obscures much more dramatic variation over a finer scale: 88 hotspots of recombination were identified (Fig. 8; see also Supplementary Fig. 7)—that is, one per 57 kb—with hotspots detected in each of the ten regions (from 4 in 12q12 to 14 in 2q37.1). Across the 5 Mb, we estimate that about 80% of all recombination has taken place in about 15% of the sequence (Fig. 9, see also refs 46, 49).

A block-like structure of human LD. With most human recombination occurring in recombination hotspots, the breakdown of LD is often discontinuous. A 'block-like' structure of LD is visually apparent in Fig. 8 and Supplementary Fig. 7: segments of consistently high D' that break down where high recombination rates, recombination hotspots and obligate recombination events⁵⁰ all cluster.

When haplotype blocks are more formally defined in the ENCODE data (using a method based on a composite of local D'

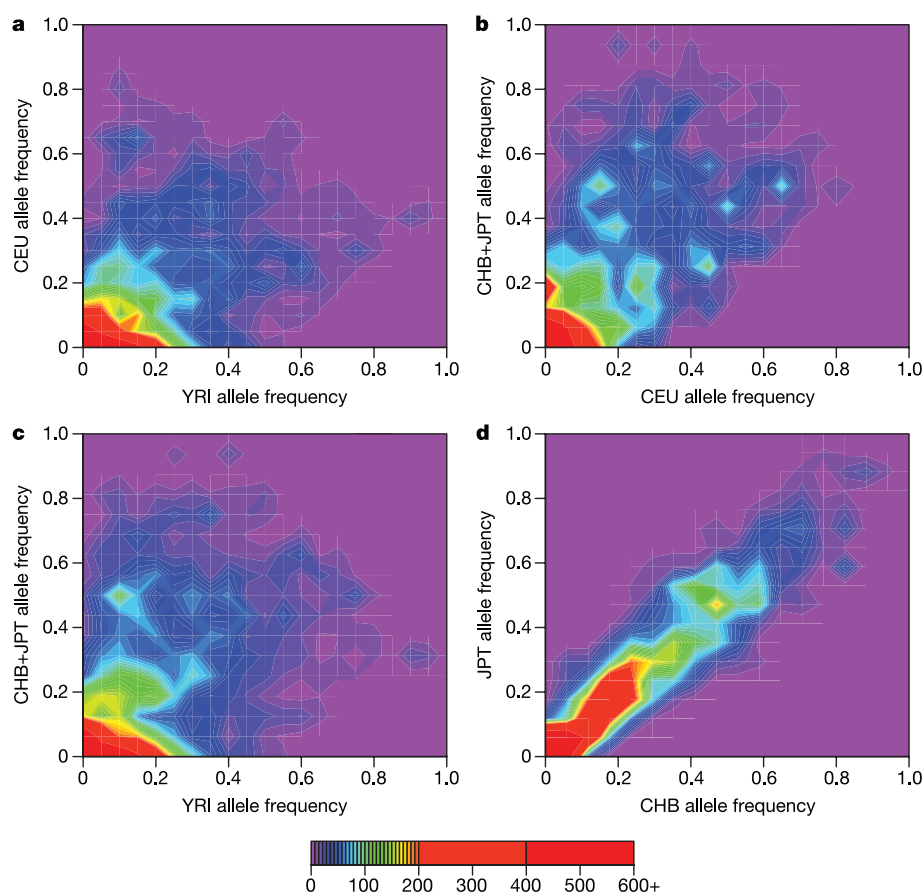


Figure 6 | Comparison of allele frequencies in the ENCODE data for all pairs of analysis panels and between the CHB and JPT sample sets. For each polymorphic SNP we identified the minor allele across all panels (a–d) and then calculated the frequency of this allele in each analysis panel/sample set. The colour in each bin represents the number of SNPs that display each

given set of allele frequencies. The purple regions show that very few SNPs are common in one panel but rare in another. The red regions show that there are many SNPs that have similar low frequencies in each pair of analysis panels/sample sets.

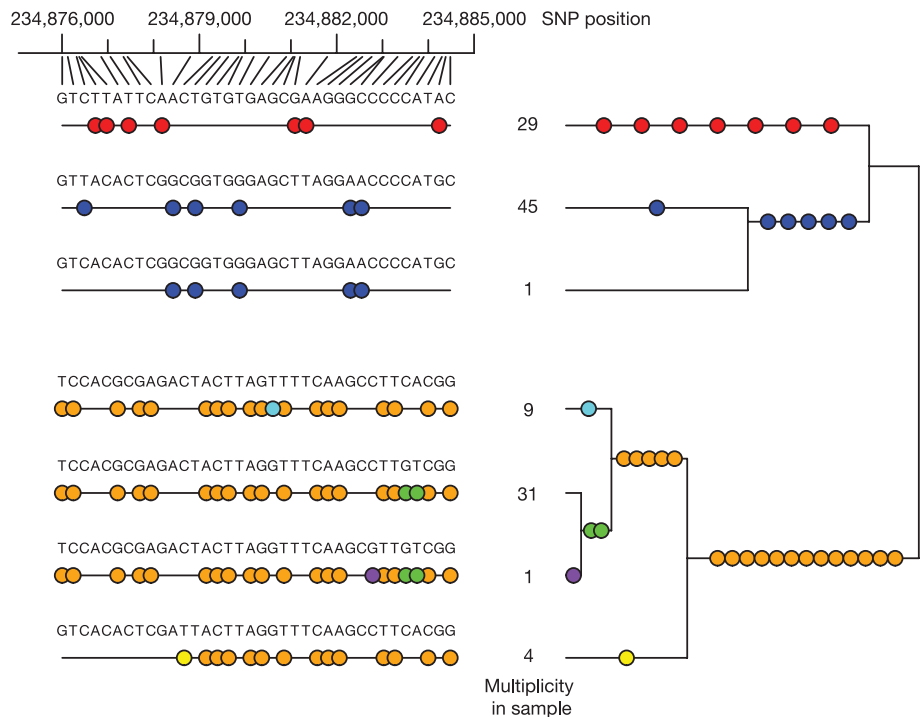


Figure 7 | Genealogical relationships among haplotypes and r^2 values in a region without obligate recombination events. The region of chromosome 2 (234,876,004–234,884,481 bp; NCBI build 34) within ENr131.2q37 contains 36 SNPs, with zero obligate recombination events in the CEU samples. The left part of the plot shows the seven different haplotypes observed over this region (alleles are indicated only at SNPs), with their respective counts in the data. Underneath each of these haplotypes is a

binary representation of the same data, with coloured circles at SNP positions where a haplotype has the less common allele at that site. Groups of SNPs all captured by a single tag SNP (with $r^2 \geq 0.8$) using a pairwise tagging algorithm^{53,54} have the same colour. Seven tag SNPs corresponding to the seven different colours capture all the SNPs in this region. On the right these SNPs are mapped to the genealogical tree relating the seven haplotypes for the data in this region.

values³⁰, or another based on the four gamete test⁵¹), most of the sequence falls into long segments of strong LD that contain many SNPs and yet display limited haplotype diversity (Table 5).

Specifically, addressing concerns that blocks might be an artefact of low marker density⁵², in these nearly complete data most of the sequence falls into blocks of four or more SNPs (67% in YRI to 87% in CEU) and the average sizes of such blocks are similar to initial estimates³⁰. Although the average block spans many SNPs (30–70), the average number of common haplotypes in each block ranged only from 4.0 (CHB + JPT) to 5.6 (YRI), with nearly all haplotypes in each block matching one of these few common haplotypes. These results confirm the generality of inferences drawn from disease-mapping studies²⁷ and genomic surveys with smaller sample sizes²⁹ and less complete data³⁰.

Long-range haplotypes and local patterns of recombination. Although haplotypes often break at recombination hotspots (and block boundaries), this tendency is not invariant. We identified all

unique haplotypes with frequency more than 0.05 across the 269 individuals in the phased data, and compared them to the fine-scale recombination map. Figure 10 shows a region of chromosome 19 over which many such haplotypes break at identified recombination hotspots, but others continue. Thus, the tendency towards co-localization of recombination sites does not imply that all haplotypes break at each recombination site.

Some regions display remarkably extended haplotype structure based on a lack of recombination (Supplementary Fig. 8a, b). Most striking, if unsurprising, are centromeric regions, which lack recombination: haplotypes defined by more than 100 SNPs span several megabases across the centromeres. The X chromosome has multiple regions with very extensive haplotypes, whereas other chromosomes typically have a few such domains.

Most global measures of LD become more consistent when measured in genetic rather than physical distance. For example, when plotted against physical distance, the extent of pairwise LD

Table 5 | Haplotype blocks in ENCODE regions, according to two methods

Parameter	YRI	CEU	CHB + JPT
Method based on a composite of local D' values ³⁰			
Average number of SNPs per block	30.3	70.1	54.4
Average length per block (kb)	7.3	16.3	13.2
Fraction of genome spanned by blocks (%)	67	87	81
Average number of haplotypes (MAF ≥ 0.05) per block	5.57	4.66	4.01
Fraction of chromosomes due to haplotypes with MAF ≥ 0.05 (%)	94	93	95
Method based on the four gamete test ⁵¹			
Average number of SNPs per block	19.9	24.3	24.3
Average length per block (kb)	4.8	5.9	5.9
Fraction of genome spanned by blocks (%)	86	84	84
Average number of haplotypes (MAF ≥ 0.05) per block	5.12	3.63	3.63
Fraction of chromosomes due to haplotypes with MAF ≥ 0.05 (%)	91	95	95

varies by chromosome; when plotted against average recombination rate on each chromosome (estimated from pedigree-based genetic maps) these differences largely disappear (Supplementary Fig. 6). Similarly, the distribution of haplotype length across chromosomes is less variable when measured in genetic rather than physical distance. For example, the median length of haplotypes is 54.4 kb on chromosome 1 compared to 34.8 kb on chromosome 21. When measured in genetic distance, however, haplotype length is much more similar: 0.104 cM on chromosome 1 compared to 0.111 cM on chromosome 21 (Supplementary Fig. 9).

The exception is again the X chromosome, which has more extensive haplotype structure after accounting for recombination rate (median haplotype length = 0.135 cM). Multiple factors could

explain different patterns on the X chromosome: lower SNP density, smaller sample size, restriction of recombination to females and lower effective population size.

A view of LD focused on the putative causal SNP

Although genealogy and recombination provide insight into why nearby SNPs are often correlated, it is the redundancies among SNPs that are of central importance for the design and analysis of association studies. A truly comprehensive genetic association study must consider all putative causal alleles and test each for its potential role in disease. If a causal variant is not directly tested in the disease sample, its effect can nonetheless be indirectly tested if it is correlated with a SNP or haplotype that has been directly tested.

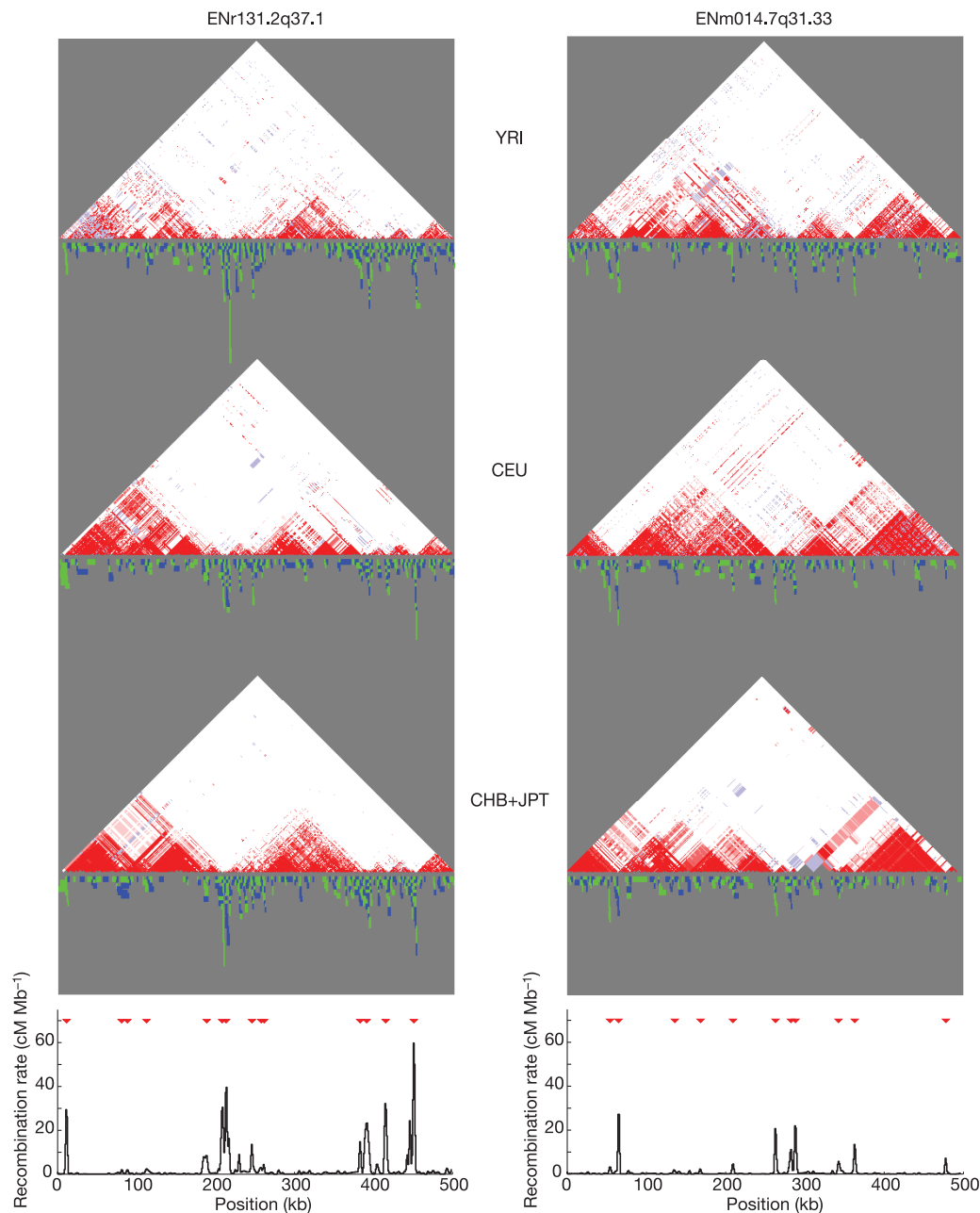


Figure 8 | Comparison of linkage disequilibrium and recombination for two ENCODE regions. For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB+JPT analysis panels are shown: white, $D' < 1$ and $\text{LOD} < 2$; blue, $D' = 1$ and $\text{LOD} < 2$; pink, $D' < 1$ and $\text{LOD} \geq 2$; red, $D' = 1$ and $\text{LOD} \geq 2$. Below each of these plots is shown the

intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles⁴⁶.

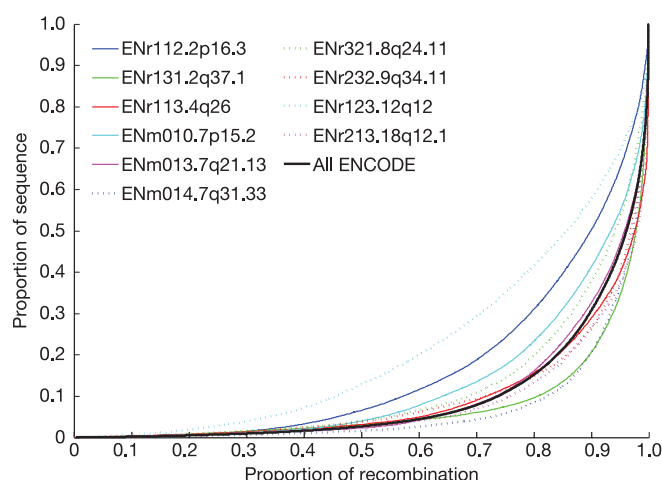


Figure 9 | The distribution of recombination events over the ENCODE regions. Proportion of sequence containing a given fraction of all recombination for the ten ENCODE regions (coloured lines) and combined (black line). For each line, SNP intervals are placed in decreasing order of estimated recombination rate⁴⁶, combined across analysis panels, and the cumulative recombination fraction is plotted against the cumulative proportion of sequence. If recombination rates were constant, each line would lie exactly along the diagonal, and so lines further to the right reveal the fraction of regions where recombination is more strongly locally concentrated.

The typical SNP is highly correlated with many of its neighbours. The ENCODE data reveal that SNPs are typically perfectly correlated to several nearby SNPs, and partially correlated to many others.

We use the term proxy to mean a SNP that shows a strong

correlation with one or more others. When two variants are perfectly correlated, testing one is exactly equivalent to testing the other; we refer to such collections of SNPs (with pairwise $r^2 = 1.0$ in the HapMap samples) as 'perfect proxy sets'.

Considering only common SNPs (the target of study for the HapMap Project) in CEU in the ENCODE data, one in five SNPs has 20 or more perfect proxies, and three in five have five or more. In contrast, one in five has no perfect proxies. As expected, perfect proxy sets are smaller in YRI, with twice as many SNPs (two in five) having no perfect proxy, and a quarter as many (5%) having 20 or more (Figs 11 and 12). These patterns are largely consistent across the range of frequencies studied by the project, with a trend towards fewer proxies at MAF < 0.10 (Fig. 11). Put another way, the average common SNP in ENCODE is perfectly redundant with three other SNPs in the YRI samples, and nine to ten other SNPs in the other sample sets (Fig. 13).

Of course, to be detected through LD in an association study, correlation need not be complete between the genotyped SNP and the causal variant. For example, under a multiplicative disease model and a single-locus χ^2 test, the sample size required to detect association to an allele scales as $1/r^2$. That is, if the causal SNP has an $r^2 = 0.5$ to one tested in the disease study, full power can be maintained if the sample size is doubled.

The number of SNPs showing such substantial but incomplete correlation is much larger. For example, using a looser threshold for declaring correlation ($r^2 \geq 0.5$), the average number of proxies found for a common SNP in CHB+JPT is 43, and the average in YRI is 16 (Fig. 12). These partial correlations can be exploited through haplotype analysis to increase power to detect putative causal alleles, as discussed below.

Evaluating performance of the Phase I map. To estimate the proportion of all common SNPs captured by the Phase I map, we

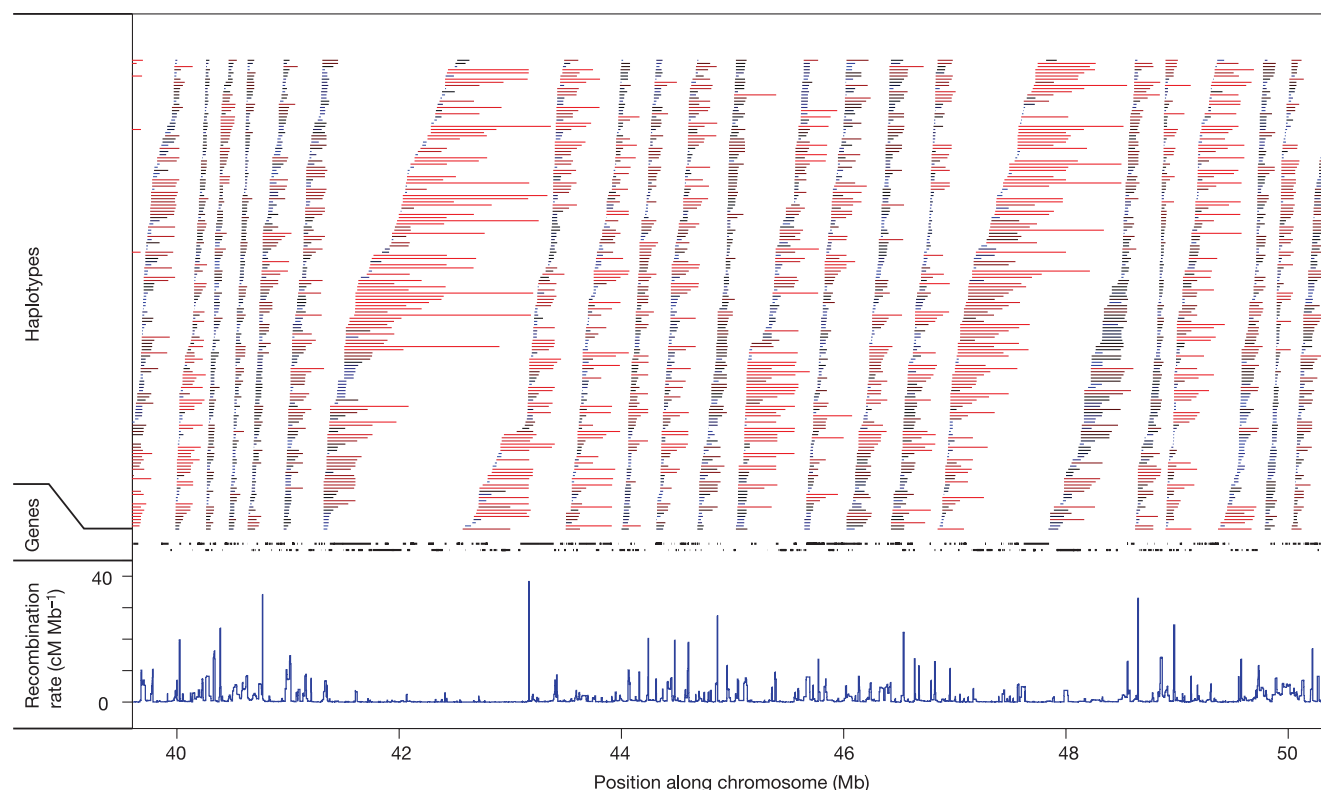


Figure 10 | The relationship among recombination rates, haplotype lengths and gene locations. Recombination rates in cM Mb^{-1} (blue). Non-redundant haplotypes with frequency of at least 5% in the combined sample (bars) and genes (black segments) are shown in an example gene-dense

region of chromosome 19 (19q13). Haplotypes are coloured by the number of detectable recombination events they span, with red indicating many events and blue few.

evaluated redundancy among SNPs on the genome-wide map, and performed simulations based on the more complete ENCODE data. The two methods give highly similar answers, and indicate that Phase I should provide excellent power for CEU, CHB and JPT, and substantial power for YRI. Phase II, moreover, will provide nearly complete power for all three analysis panels.

Redundancies among SNPs in Phase I HapMap. Redundancy offers one measure that Phase I has sampled densely in comparison to the underlying scale of correlation. Specifically, 50% (YRI) to 75% (CHB+JPT, CEU) of all SNPs on the Phase I map are highly correlated ($r^2 \geq 0.8$) to one or more others on the map (Fig. 13; see also Supplementary Fig. 10). Over 90% of all SNPs on the map have highly statistically significant correlation to one or more neighbours. These partial correlations can be combined to form haplotypes that are even better proxies for a SNP of interest.

Modelling Phase I HapMap from complete ENCODE data. A second approach to evaluating the completeness of the Phase I data involves thinning the more complete ENCODE data to match Phase I for allele frequency and SNP density. Simulated Phase I HapMaps were used to evaluate coverage in relation to the full set of common SNPs (Table 6), and provided nearly identical estimates to those above: 45% (YRI) to 74% (CHB+JPT, CEU) of all common SNPs are predicted to have a proxy with $r^2 \geq 0.8$ to a SNP included in the Phase I HapMap (Supplementary Fig. 11).

Statistical power in association studies may be more closely approximated by the average (maximal) correlation value between a SNP and its best proxy on the map, rather than by the proportion exceeding an arbitrary (and stringent) threshold. The average values for maximal r^2 to a nearby SNP range from 0.67 (YRI) to 0.85 (CEU and CHB+JPT).

Modelling Phase II HapMap from complete ENCODE data. A similar procedure was used to generate simulated Phase II HapMaps from ENCODE data (Table 6). Phase II is predicted to capture the majority of common variation in YRI: 81% of all common SNPs should have a near perfect proxy ($r^2 \geq 0.8$) to a SNP on the map, with the mean maximal r^2 value of 0.90. Unsurprisingly, the CEU, CHB and JPT samples, already well served by Phase I, are nearly perfectly captured: 94% of all common sites have a proxy on the map with $r^2 \geq 0.8$, with an average maximal r^2 value of 0.97.

These analyses indicate that the Phase I and Phase II HapMap resources should provide excellent coverage for common variation in these population samples.

Selection of tag SNPs for association studies

A major impetus for developing the HapMap was to guide the design and prioritization of SNP genotyping assays for disease association studies. We refer to the set of SNPs genotyped in a disease study as tags. A given set of tags can be analysed for association with a

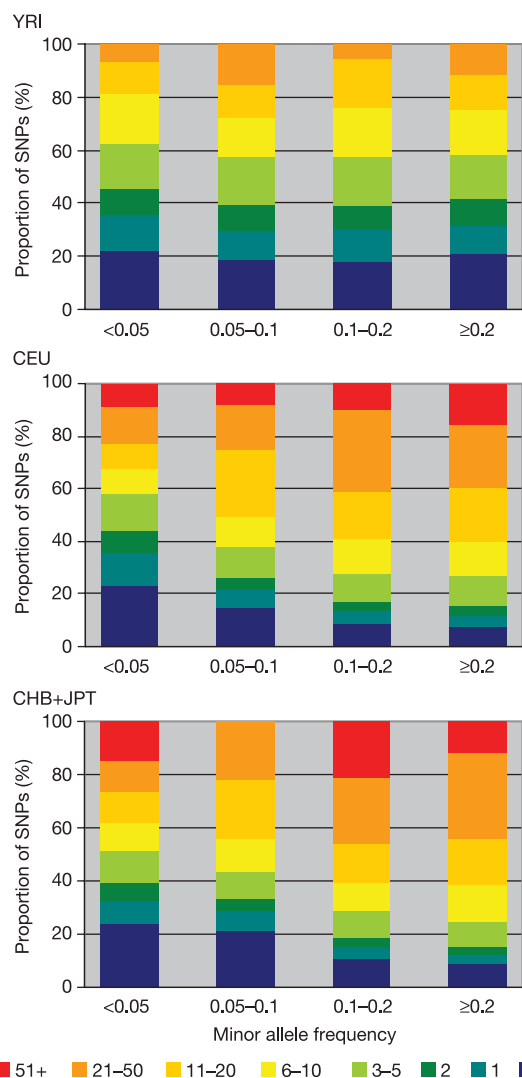


Figure 11 | The number of proxy SNPs ($r^2 \geq 0.8$) as a function of MAF in the ENCODE data.

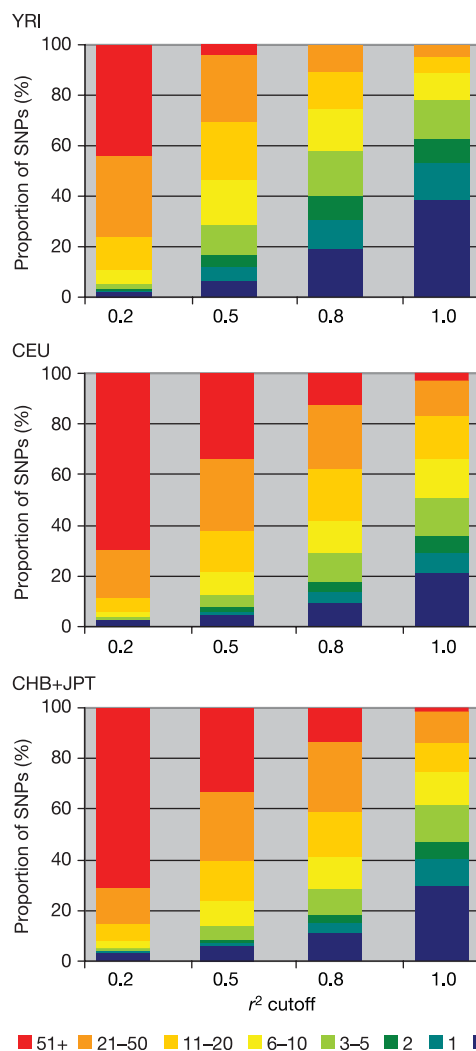


Figure 12 | The number of proxies per SNP in the ENCODE data as a function of the threshold for correlation (r^2).

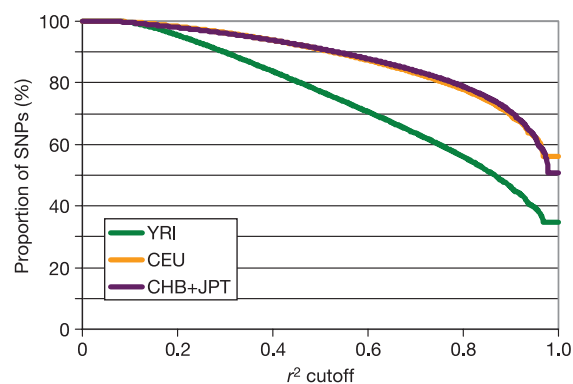


Figure 13 | Relationship in the Phase I HapMap between the threshold for declaring correlation between proxies and the proportion of all SNPs captured.

Table 6 | Coverage of simulated Phase I and Phase II HapMap to capture all common SNPs in the ten ENCODE regions

Analysis panel	Per cent maximum $r^2 \geq 0.8$	Mean maximum r^2
Phase I HapMap		
YRI	45	0.67
CEU	74	0.85
CHB+JPT	72	0.83
Phase II HapMap		
YRI	81	0.90
CEU	94	0.97
CHB+JPT	94	0.97

Simulated Phase I HapMaps were generated from the phased ENCODE data (release 16c1) by randomly picking SNPs that appear in dbSNP build 121 (excluding 'non-rs' SNPs in release 16a) for every 5-kb bin until a common SNP was picked (allowing up to three attempts per bin). The Phase II HapMap was simulated by picking SNPs at random to achieve an overall density of 1 SNP per 1 kb. These numbers are averages over 20 independent iterations for all ENCODE regions in all three analysis panels.

phenotype using a variety of statistical methods which we term tests, based either on the genotypes of single SNPs or combinations of multiple SNPs.

The shared goal of all tag selection methods is to exploit redundancy among SNPs, maximizing efficiency in the laboratory while minimizing loss of information^{24,27}. This literature is extensive and varied, despite its youth. Some methods require that a single SNP serve as a proxy for other, untyped variants, whereas other methods allow combinations of alleles (haplotypes) to serve as proxies; some make explicit use of LD blocks whereas others are agnostic to such descriptions. Although it is not practical to implement all such methods at the project website, the HapMap genotypes are freely available and investigators can apply their method of choice to the data. To assist users, both a single-marker tagging method and a more efficient multimarker method have been implemented at <http://www.hapmap.org>.

Tagging using a simple pairwise method. To illustrate general principles of tagging, we first applied a simple and widely used pairwise algorithm^{53,54}. SNPs are selected for genotyping until all common SNPs are highly correlated ($r^2 \geq 0.8$) to one or more members of the tag set.

Starting from the substantially complete ENCODE data, the density of common SNPs can be reduced by 75–90% with essentially no loss of information (Fig. 14). That is, the genotyping burden can be reduced from one common SNP every 500 bp to one SNP every 2 kb (YRI) to 5 kb (CEU and CHB+JPT). Because LD often extends for long distances, studies of short gene segments tend to underestimate the redundancy across the genome⁴³.

Although tags selected based on LD offer the greatest improvements in efficiency and information capture, even randomly chosen subsets of SNPs offer considerable efficiencies (Fig. 14).

The data also reveal a rule of diminishing returns: a small set of highly informative tags captures a large fraction of all variation, with additional tags each capturing only one or a few proxies. For example, in CHB+JPT the most informative 1% of all SNPs (one per 50 kb) is able to proxy (at $r^2 \geq 0.8$) for 40% of all common SNPs, whereas a substantial proportion of SNPs have no proxies at all.

These observations are encouraging with respect to genome-wide association studies. A set of SNPs typed every 5–10 kb across the genome (within the range of current technology) can capture nearly all common variation in the genome in the CEU and CHB+JPT samples, with more SNPs required in the YRI samples.

Tagging from the genome-wide map. Whereas analysis of the complete ENCODE data set reveals the maximal efficiency likely to be possible with this tag selection strategy, analysis of the Phase I map illuminates the extent to which the current resource can be used for near-term studies. Specifically, using the same pairwise tagging approach above, 260,000 (CHB+JPT) to 474,000 (YRI) SNPs are required to capture all common SNPs in the Phase I data set (Table 7). That is, being incomplete and thus less redundant, the Phase I data are much less compressible by tag SNP selection than are the ENCODE data. Nevertheless, even at this level a half to a third of all SNPs can be selected as proxies for the remainder (and, by inference, the bulk of other common SNPs in the genome).

Increasing the efficiency of tag SNPs. Although the pairwise method

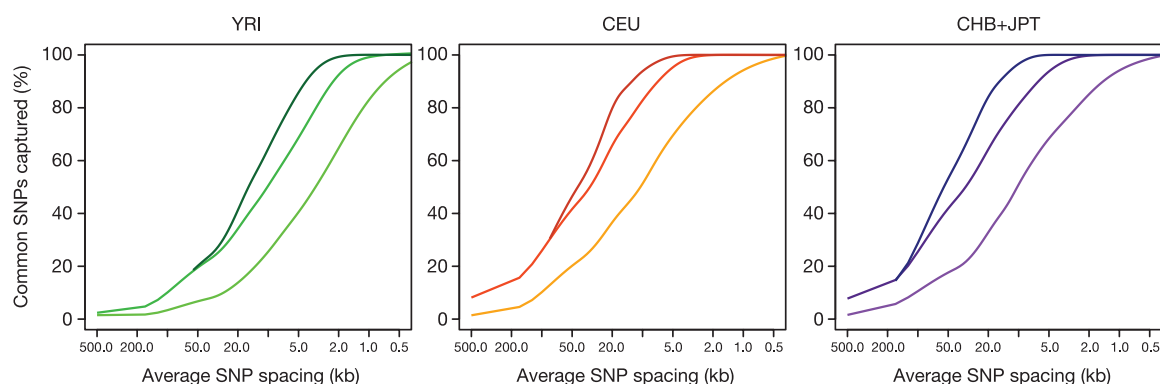


Figure 14 | Tag SNP information capture. The proportion of common SNPs captured with $r^2 \geq 0.8$ as a function of the average tag SNP spacing is shown for the phased ENCODE data, plotted (left to right) for tag SNPs prioritized

by Tagger (multimarker and pairwise) and for tag SNPs picked at random. Results were averaged over all the ENCODE regions.

is simple, complete and straightforward, efficiency can be improved with a number of simple changes. First, relaxing the threshold on r^2 for tag SNP selection substantially reduces the number of tag SNPs selected, with only a modest decrease in the correlations among SNPs (Table 7). For example, reducing the r^2 threshold from 0.8 to 0.5 decreases the number of tag SNPs selected from the HapMap by 39% in CHB+JPT (260,000 to 159,000) and 32% in YRI (474,000 to 325,000). The average r^2 value between tags and other (unselected) SNPs falls much less dramatically than the number of tags selected, increasing efficiency. Whether such a loss of power is justified by the disproportionate reduction in work is a choice each investigator will need to make.

A second enhancement exploits multimarker haplotypes. Many investigators have discussed using multiple SNPs (in haplotypes and regression models) to serve as proxies for untyped sites^{55–58}, which may reduce the number of tags required and increase the power of analyses performed. Figure 14 illustrates the point with one such method⁵⁵, showing that a multimarker method allows greater coverage for a fixed set of markers (or, alternatively, fewer markers to achieve the same coverage). Although a full consideration of this issue is beyond the scope of this paper, the availability of these and other data should allow the comparison and application of such methods.

A third approach to increasing efficiency is to prioritize tags based on the number of other SNPs captured. Whereas 260,000 SNPs are required to provide $r^2 \geq 0.8$ for all SNPs in the Phase I HapMap (CHB+JPT), the best 10,000 such SNPs (4%) capture 22% of all common variable sites with $r^2 \geq 0.8$ (Table 8). Such prioritization can be applied using different weights for SNPs based on genomic annotation (for example, non-synonymous coding SNPs, SNPs in conserved non-coding sequence, and candidate genes of biological interest).

Tag transferability across populations. The most complete set of tags would be those based on all 269 samples; however, many studies may be performed in individuals more closely related to one particular HapMap population, and efficiency may be gained by selecting tags only from that population sample. (Selecting tags in a HapMap population sample that is known to be more distantly related than is another, for example, using CEU to pick tags for a study of Japanese, seems inefficient.)

An important question is how tags selected in one or more analysis panels will transfer to disease studies performed in these or other populations. Our data do not address this question directly, although the known similarity of allele and haplotype frequencies across populations within continents⁴¹ is encouraging. More data are clearly needed, however.

Tag selection based on initial genotyping. Whereas the discussions above assume *de novo* selection of SNPs, many investigators will have already performed initial studies, and wish to design follow-on experiments. The HapMap data can be used to highlight SNPs that might potentially explain a positive association signal, or those that were poorly captured (and thus still need to be tested) after a negative scan. In cases where multiple SNPs are both associated with the trait and with each other, the HapMap data can be queried to identify whether samples from any other analysis

panel show a breakdown of LD in that region, and thus the possibility of narrowing the span over which the causal variant may reside.

Applications to the analysis of association data

Beyond guiding selection of tag SNPs, HapMap data can inform the subsequent analysis and interpretation in disease association studies.

Analysis of an existing genotype data set. The HapMap can be used to inform association testing, regardless of how tags were selected. Specifically, as long as the SNPs genotyped in a disease study have also been typed in the HapMap samples, it is possible to identify which SNPs are well captured by the genotyped SNPs (either singly, or in haplotype combinations), and which are not⁵⁵.

This is of particular importance for genome-wide association studies performed using array-based, standardized genotyping reagents, which do not allow investigators to choose their own sets of tag SNPs. The Affymetrix 120K SNP array data included in Phase I of the HapMap provides a simple example: in CEU 48% of HapMap SNPs have substantial pairwise correlation ($r^2 \geq 0.5$) to one or more of the 120K SNPs on the array. An additional 13%, however, are not correlated to a single SNP, but are to a specific haplotype of two members of the 120K panel. By identifying such haplotype predictors in the HapMap, and testing them (in addition to the single SNPs) in a disease study, it is likely that power will be increased (I. Pe'er *et al.*, manuscript in preparation).

Evaluating statistical significance and interpreting results. An important challenge in genome-wide association testing is to develop statistical procedures that minimize false positives without greatly sacrificing true positives. The challenge is amplified by the correlated nature of polymorphism data, which makes simple frequentist approaches that assume independence (such as Bonferroni correction) highly conservative. To illustrate this point, we used the ENCODE data to estimate the 'effective number of independent tests' (the statistical burden of testing all common ($MAF \geq 0.05$) variation) across large genomic regions. Specifically, we re-sampled from the phased ENCODE chromosomes to create mock case-control panels in which all common SNPs were observed, but there was not a causal allele. The resulting χ^2 distribution for association indicates that complete testing of common variation in each 500-kb region is equivalent to performing about 150 independent statistical tests (in CEU and CHB+JPT) and about 350 tests (in YRI). Although it will probably be desirable to perform such empirical estimates of significance within each disease study, these results illustrate how Bonferroni correction overestimates the statistical penalty of performing many correlated tests.

Study of less common alleles. We have focused primarily on the hypothesis that a single, common causal allele exists, and needs to be tested for association to disease. Of course, in many cases the causal allele(s) will be less common, and might be missed by such an approach.

It is possible to perform additional haplotype tests, beyond those

Table 7 | Number of selected tag SNPs to capture all observed common SNPs in the Phase I HapMap

r^2 threshold*	YRI	CEU	CHB + JPT
$r^2 \geq 0.5$	324,865	178,501	159,029
$r^2 \geq 0.8$	474,409	293,835	259,779
$r^2 = 1.0$	604,886	447,579	434,476

Tag SNPs were picked to capture common SNPs in HapMap release 16c1 using the software program Haploview.

*Pairwise tagging at different r^2 thresholds.

Table 8 | Proportion of common SNPs in Phase I captured by sets of tag SNPs

Tag SNP set size	Common SNPs captured (%)		
	YRI	CEU	CHB + JPT
10,000	12.3	20.4	21.9
20,000	19.1	30.9	33.2
50,000	32.7	50.4	53.6
100,000	47.2	68.5	72.2
250,000	70.1	94.1	98.5

As in Table 7, tag SNPs were picked to capture common SNPs in HapMap release 16c1 using Haploview, selecting SNPs in order of the fraction of sites captured. Common SNPs were captured by fixed-size sets of pairwise tags at $r^2 \geq 0.8$.

that capture known polymorphisms, in the hope of capturing less common or unrepresented alleles⁵⁶. Such haplotype analysis has a long history and proven value in mendelian genetics; the causal mutation is generally rare and unexamined during initial genotyping, but is frequently recognized by its presence on a long, unique haplotype of common alleles^{18,19,59–62}.

Admixture mapping. Although not designed specifically to enable admixture mapping⁶³, the HapMap has helped lay the groundwork for this approach. Admixture mapping requires a map of SNPs that are highly differentiated in frequency across population groups. By typing many SNPs in samples from multiple geographical regions, the data have helped to identify such SNPs for the design of genome-wide admixture mapping panels^{64,65} and can be further used to identify candidate SNPs with large allele frequency differences for follow-up of positive admixture scan results⁶⁶.

Loss of heterozygosity in tumours. Loss of heterozygosity (LOH) in tumour tissue can be a powerful indicator of the location of tumour suppressor genes, and genome-wide, fine-scale LOH analysis has been empowered by genome-wide SNP arrays⁶⁷. Germline DNA is not always available from the same subjects, however, and even if available, typing of germline DNA doubles project costs. In lower density scans for LOH (with markers far apart relative to the scale of LD), long runs of homozygosity in tumours are nearly always indicative of LOH. However, at higher densities runs of homozygosity can be due to haplotype homozygosity in the inherited germline DNA, rather than LOH.

The HapMap data can help minimize this difficulty; previous probabilities for homozygosity based on known frequencies of haplotypes in the HapMap data can be used to distinguish homozygosity due to haplotype sharing rather than LOH⁶⁸.

Identifying structural variants in HapMap data

Structural variations—segments where DNA is deleted, duplicated, or rearranged—are common^{69,70} and have an important role in diseases^{71–73}. The HapMap can provide some insight into structural variation because, in many cases, structural variants reveal themselves through signatures in SNP genotype data. In particular, polymorphic deletions are important to discover, because loss of genetic material is of obvious functional relevance, and results in aberrant patterns of SNP genotypes. These include apparent non-mendelian inheritance of SNP alleles, null genotypes and deviations from Hardy–Weinberg equilibrium. However, such SNPs are routinely discarded as technical failures of genotyping.

Thus, we scanned the unfiltered Phase I HapMap data using an approach developed and validated to identify polymorphic deletions from clusters of SNPs with aberrant genotype patterns (calibrated across the multiple centres and genotyping platforms⁷⁴). In total, 541 candidate deletion polymorphisms were identified, of which 150 were common enough to be observed as homozygotes.

The properties of these candidate deletions, including experimental validation of 90 candidates, are described in ref. 74. Validated

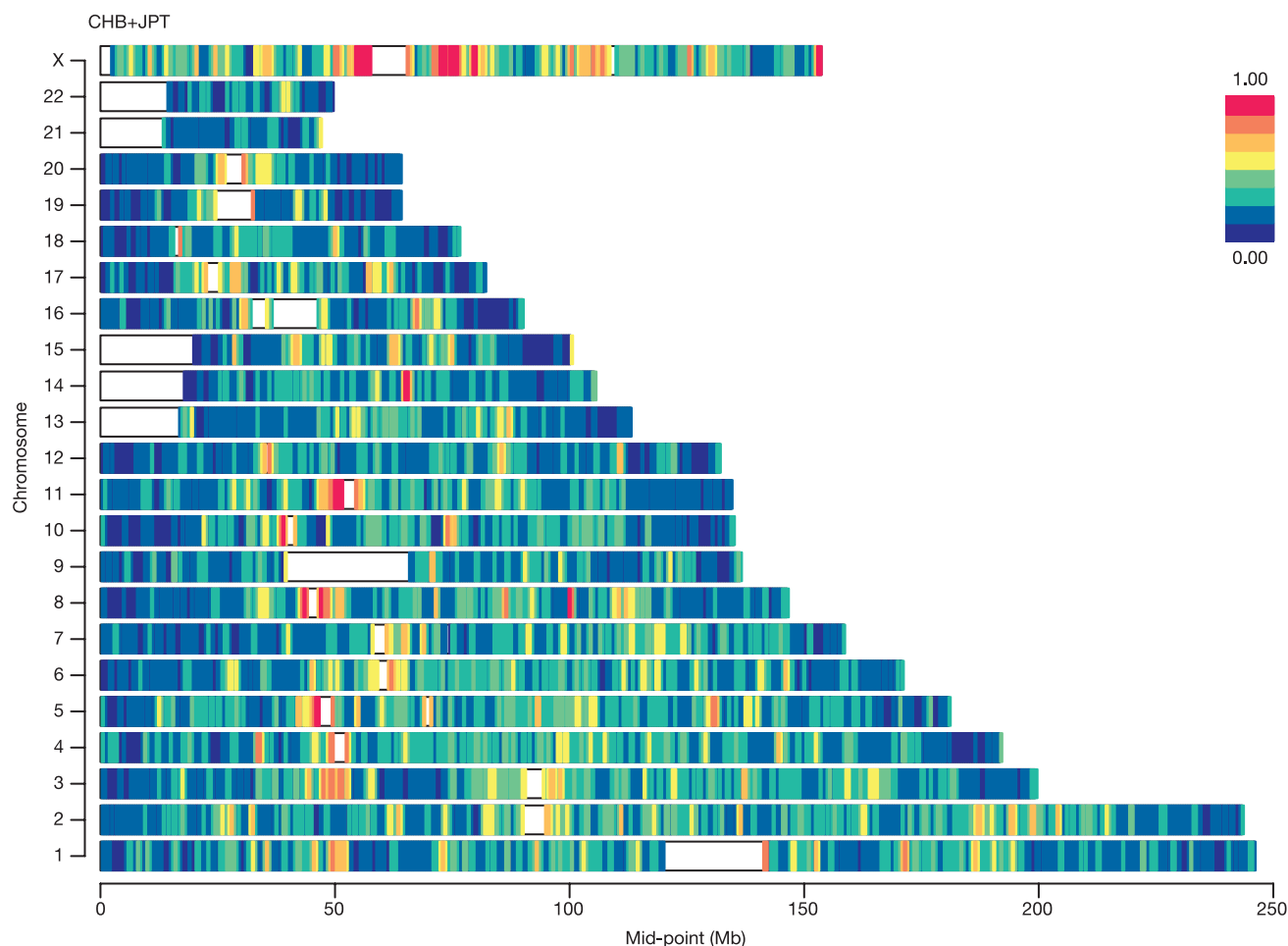


Figure 15 | Length of LD spans. We fitted a simple model for the decay of linkage disequilibrium¹⁰³ to windows of 1 million bases distributed throughout the genome. The results of model fitting are summarized for the

CHB+JPT analysis panel, by plotting the fitted r^2 value for SNPs separated by 30 kb. The overall pattern of variation was very similar in the other analysis panels⁸⁴ (see Supplementary Information).

polymorphisms include 10 that remove coding exons of genes, such that in many cases individuals are homozygous null for the encoded transcript. Analysis of confirmed deletions often shows strong LD with nearby SNPs, indicating that LD-based approaches can be useful for detecting disease associations due to structural (as well as SNP) variants.

Polymorphic inversions may also be reflected in the HapMap data as long regions where multiple SNPs are perfectly correlated: because recombination between an inverted and non-inverted copy is lethal, the inverted and non-inverted copies of the region evolve independently. A striking example corresponds to the known inversion polymorphism on chromosome 17, present in 20% of the CEU chromosomes, that has been associated with fertility and total recombination rate in females among Icelanders⁷⁵. Long LD may also arise, however, due to a low recombination rate or certain forms of natural selection, as discussed below.

Insights into recombination and natural selection

In addition to its intended function as a resource for disease studies, the HapMap data provide clues about the biology of recombination and history of natural selection.

A genome-wide map of recombination rates at a fine scale. On the basis of the HapMap data, we created a fine-scale genetic map spanning the human genome (Supplementary Fig. 12), including 21,617 identified recombination hotspots (one per 122 kb).

Both the number and intensity of hotspots contribute to overall variation in recombination rate. For example, we selected 25 regions of 5 Mb as having the highest ($>2.75 \text{ cM Mb}^{-1}$) and lowest ($<0.5 \text{ cM Mb}^{-1}$) rates of recombination in the deCODE (pedigree-based) genetic map⁴⁸. We detected recombination hotspots in all regions, even the lowest. But in the high cM Mb^{-1} regions hotspots are more closely spaced (one per 84 kb) and have a higher average

intensity (0.124 cM) as compared to the low cM Mb^{-1} regions (one every 208 kb, and 0.051 cM , respectively).

Estimates of recombination rates and identified hotspots are robust to the specific markers and samples studied. Specifically, we compared these results to a similar analysis⁷⁶ of the data of ref. 77 (with about 1.6 million SNPs genotyped in 71 individuals). We find nearly complete correlation in rate estimates at a coarse scale (5 Mb) between these two surveys ($r^2 = 0.99$) and to the pedigree map ($r^2 = 0.95$). Very substantial correlation is found at finer scales: $r^2 = 0.8$ at 50 kb and $r^2 = 0.59$ at 5 kb. Moreover, of the 21,617 hotspots identified using the HapMap data, 78% (16,923) were also identified using the data of ref. 77.

The ability to detect events depends on marker density, with the larger number of SNPs studied by ref. 77 increasing power to detect hotspots, and presumably precision of rate estimates. There are, however, substantial genomic regions where the HapMap data have a higher SNP density. For example, more hotspots are detected on chromosomes 9 and 19 from the HapMap data. We expect that Phase II of HapMap will provide a genome-wide recombination map of substantially greater precision than either ref. 77, or Phase I, at fine scales.

Little is yet known about the molecular determinants of recombination hotspots. In an analysis of the data of ref. 77, another study (ref. 76) found significant evidence for an excess of the THE1A/B retrotransposon-like elements within recombination hotspots, and more strikingly for a sixfold increase of a particular motif (CCTCCCT) within copies of the element in hotspots, compared to copies of the element outside hotspots. In analysing the HapMap data, we confirmed these findings (Supplementary Fig. 13). Furthermore, THE1B elements with the motif are particularly enriched within 1.5 kb of the centre of the hotspots compared to flanking sequence ($P < 10^{-16}$).

Correlations of LD with genomic features. Variation in recombination rate is important, in large part, because of its impact on LD. We thus examined genome-wide LD for correlation to recombination rates, sequence composition and gene features.

We confirmed previous observations that LD is generally low near telomeres, elevated near centromeres, and correlated with chromosome length (Fig. 15; see also Supplementary Figs 8b and 14)^{48,78–80}. These patterns are due to recombination rate variation as discussed above. We also confirmed previously described relationships between LD and G+C content^{78,81,82}, sequence polymorphism⁸³ and repeat composition^{78,82}.

We observe, for the first time, that LD tracks with both the density and functional classification of genes. We examined quartiles of the genome based on extent of LD, and looked for correlations to gene density. Surprisingly, we find that both the top and bottom quartiles of the genome have greater gene density as compared to the middle quartiles (6.7 as compared to 6.1 genes per Mb), as well as percentage of bases in codons (1.24% as compared to 1.08%). We have no explanation for this observation.

Although the majority of gene classes are equally divided between these two extreme quartiles of the genome, some classes of genes show a marked skew in their distribution^{64,84,85}. Genes involved in immune responses and neurophysiological processes are more often located in regions of low LD, whereas genes involved in DNA and RNA metabolism, response to DNA damage and the cell cycle are preferentially located in regions of strong linkage disequilibrium. It is intriguing to speculate that the extent of LD (and sequence diversity) might track with gene function due to natural selection, with increased diversity being favoured in genes involved in interface with the environment such as the immune response⁸⁶, and disadvantageous for core cell biological processes such as DNA repair and packaging^{87,88}.

Natural selection. The preceding observation highlights the hypothesis that signatures of natural selection are present in the HapMap data. The availability of genome-wide variation data makes it

Table 9 | High-differentiation non-synonymous SNPs

Chromosome	Position (base number)	Gene*	SNP
1	54,772,383	THEA	rs1702003
1	156,000,000	FY	rs12075
1	244,000,000	Q8NGY8_human†	rs7555046
2	3,184,917	COLEC11	rs7567833
2	73,563,622	ALMS1	rs3813227
2	73,589,553	ALMS1	rs6546837
2	73,591,645	ALMS1	rs6724782
2	73,592,163	ALMS1	rs6546839
2	73,629,222	ALMS1	rs2056486
2	73,629,311	ALMS1	rs10193972
2	109,000,000	EDAR	rs3827760
3	182,000,000	FXR1	rs11499
3	185,000,000	MCF2L2	rs7639705
4	41,844,599	SLC30A9	rs1047626
4	46,567,077	ENSG00000172895.1	rs5825
4	101,000,000	ADH1B	rs1229984
8	10,517,787	RPL1	rs6601495
8	146,000,000	SLC39A4	rs1871534
10	50,402,145	ERCC6	rs4253047
10	71,002,210	NEUROG3	rs4536103
11	46,701,579	F2	rs5896
15	46,213,776	SLC24A5	rs1426654
15	61,724,262	HERC1	rs7162473
16	30,996,126	ZNF646	rs749670
16	46,815,699	ABCC11	rs17822931
17	26,322,430	RNF135	rs7225888
17	26,399,303	ENSG00000184253.2	rs6505228
18	66,022,323	RTTN	rs3911730
19	5,782,891	FUT6	rs364637
19	47,723,209	CEACAM1	rs8110904
22	18,164,095	GNB1L	rs2073770
X	65,608,007	EDA2R	rs1385699

* Where no standard gene abbreviation exists, the ENSEMBL gene ID has been given.

† It is unclear from current annotations whether this is a pseudogene.

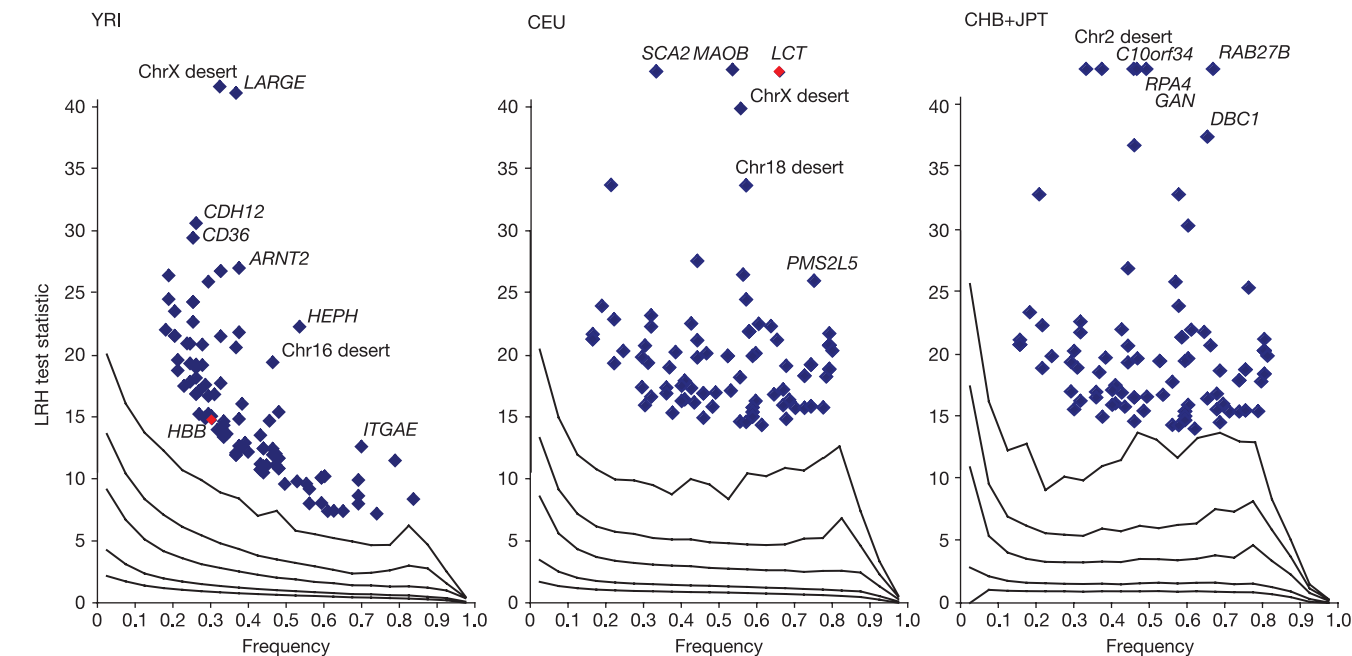


Figure 16 | The distribution of the long range haplotype (LRH⁹²) test statistic for natural selection. In the YRI analysis panel, diversity around the *HBB* gene is highlighted by the red point. In the CEU analysis panel, diversity within the *LCT* gene region is similarly highlighted.

possible to scan the genome for such signatures to discover genes that were subject to selection during human evolution⁸⁹; the HapMap data also provide a genome-wide empirical distribution against which previous claims of selection can be evaluated (rather than relying solely on theoretical computer simulations).

Natural selection influences patterns of genetic variation in various ways, such as through the removal of deleterious mutations, the fixation of advantageous variants, and the maintenance of multiple alleles through balancing selection. Each form of selection may have occurred uniformly across the world (and thus be represented in all human populations) or have been geographically localized (and thus differ among populations).

Nearly all methods for recognizing natural selection rely on the collection of complete sequence data. The HapMap Project's focus on common variation—and the process of SNP selection that achieved a preponderance of high-frequency alleles (Fig. 5)—thus prevents their straightforward application. Adjusting for the effect of SNP choice is

complex, moreover, because SNP choice varied over time as dbSNP evolved, and was implemented locally at each centre.

For these reasons, we focus here on two types of analysis. First, we examined the distributions of signatures of selection across the genome. Although the absolute value of these measures is difficult to interpret (owing to SNP ascertainment), the most extreme cases in a genome-wide distribution are important candidates to evaluate for selection. Second, we compared across functional categories, because SNP choice was largely agnostic to such features, and thus systematic differences may be a sign of selection.

The outcomes of these analyses confirm a number of previous hypotheses about selection and identify new loci as candidates for selection.

Evidence for selective sweeps in particular genomic regions. First we consider population differentiation, generally accepted as a clue to past selection in one of the populations. The HapMap data reveal 926 SNPs with allele frequencies that differ across the analysis panels

Table 10 | Candidate loci in which selection occurred

Chromosome	Position (base number) at centre	Genes in region	Population	Haplotype frequency	Empirical P-value
2	137,224,699	<i>LCT</i>	CEU	0.65	1.25×10^{-9}
5	22,296,347	<i>CDH12</i> , <i>PMCHL1</i>	YRI	0.25	5.77×10^{-8}
7	79,904,387	<i>CD36</i>	YRI	0.24	2.72×10^{-6}
7	73,747,934	<i>PMS2L5</i> , <i>WBSCR16</i>	CEU	0.76	3.37×10^{-6}
12	109,892,896	<i>CUTL2</i>	CEU	0.36	7.95×10^{-9}
15	78,558,508	<i>ARNT2</i>	YRI	0.32	6.92×10^{-7}
16	75,661,011	Desert	YRI	0.46	5.01×10^{-7}
17	3,945,580	<i>ITGAE</i> , <i>GSG2</i> , <i>HSA277841</i> , <i>CAMKK1</i> , <i>P2RX1</i>	YRI	0.70	9.26×10^{-7}
18	24,502,756	Desert	CEU	0.57	2.23×10^{-7}
22	32,459,471	<i>LARGE</i>	YRI	0.36	7.82×10^{-9}
X	20,171,291	Desert	YRI	0.33	5.02×10^{-9}
X	64,323,320	<i>HEPH</i>	YRI	0.55	3.02×10^{-8}
X	42,763,073	<i>MAOB</i>	CEU	0.53	4.21×10^{-9}
X	34,399,948	Desert	CEU	0.57	8.85×10^{-8}

in a manner as extreme as the well-accepted example of selection at the Duffy (*FY*) locus (Supplementary Fig. 8c). Of these 926 SNPs, 32 are non-synonymous coding SNPs and many others occur in transcribed regions, making them strong candidates for functional polymorphisms that have experienced geographically restricted selection pressures (see Table 9 and Supplementary Information for details). In particular, the *ALMS1* gene on chromosome 2 has six amino acid polymorphisms that show very strong population differentiation.

Another signature of an allele having risen to fixation through selection is that all other diversity in the region is eliminated (known as a selective sweep). We identified extreme outliers in the joint distribution of heterozygosity (as assessed from shotgun sequencing SNP discovery projects) and either population differentiation or skewing of allele frequency towards rare alleles in each analysis panel (Supplementary Fig. 15). We identified 19 such genomic regions (13 on autosomes, 6 on the X chromosome) as candidates for future study (Supplementary Table 4); these include candidates for population-specific sweeps and sweeps in the ancestral population. Encouragingly, this analysis includes among its top-scoring results the *LCT* gene, which influences the ability to digest dairy products⁹⁰ and has been shown to be subject to past natural selection⁹¹.

Long haplotypes as candidates for natural selection. Selective sweeps that fail to fix in the population, as well as balancing selection, lead to haplotypes that are relatively high in frequency and long in duration. In the *HLA* region (which is widely believed to have been influenced by balancing selection) multiple haplotypes of 500 SNPs that extend more than 1 cM in length are observed with a frequency in the HapMap samples of more than 1%. We identified other such occurrences of long haplotypes across the genome (Supplementary Fig. 8 and Supplementary Tables 5 and 6).

An approach to long haplotypes designed specifically to identify regions having undergone partial selective sweeps is the long range haplotype (LRH) test^{91,92}, which compares the length of each haplotype to that of others at the locus, matched across the genome based on frequency. Previously identified outliers to the genome-wide distribution for the LRH test (Fig. 16) that have been identified as candidates for selection include the *LCT* gene in the CEU sample (empirical P -value = 1.3×10^{-9}), which was an outlier for the heterozygosity/allele frequency test above, and the *HBB* gene (empirical P -value = 1.39×10^{-5}) in the YRI sample. However, most of the strongest signals in the LRH test (Table 10) were not previously hypothesized as undergoing selection.

These four tests overlap only partially in the hypotheses they address—heterozygosity, for example, is sensitive to older sweeps, whereas the haplotype tests are most powerful for partial sweeps—but encouragingly some candidate regions are found by more than one test. In particular, six regions are identified both by long haplotypes and by low heterozygosity, and three regions (*LCT* on chromosome 2, and two regions on the X chromosome at 20 and 65 Mb) are identified by three different tests.

Confirming purifying selection at conserved non-coding elements. Finally, we used the HapMap data to test an important hypothesis from comparative genomics. Genomic sequencing has shown that about 5% of the human sequence is highly conserved across species, yet less than half of this sequence spans known functional elements such as exons⁴⁵. It is widely assumed that conserved non-genic sequences lack diversity because of selective constraint (that is, purifying selection), but such regions may simply be coldspots for mutation, and thus be of little value as candidates for functional study.

Analysis of allele frequencies helps to resolve this uncertainty. Functional constraint, but not a low mutation rate, results in a downward skew in allele frequencies for conserved sequences as compared to neutral sequences^{93,94}. We find that conserved non-genic sequences display a greater skew towards rare alleles than do

intergenic regions, as predicted under purifying selection. This skew is less extreme than that observed for exons (Supplementary Fig. 16), reflecting either stronger purifying selection or the prioritization of coding SNPs for genotyping by the HapMap centres regardless of validation status. This novel evidence for ongoing constraint shows that conserved non-genic sequences are not mutational coldspots, and thus remain of high interest for functional study.

Conclusions

The International HapMap Project set out to create a resource that would accelerate the identification of genetic factors that influence medical traits. Analyses reported here confirm the generality of hotspots of recombination, long segments of strong LD, and limited haplotype diversity. Most important is the extensive redundancy among nearby SNPs, providing (1) the potential to extract extensive information about genomic variation without complete resequencing, and (2) efficiencies through selection of tag SNPs and optimized association analyses. Beyond the biomedical context, these data have made it possible to identify deletion variants in the genome, explore the nature of fine-scale recombination and identify regions that may have been subject to natural selection.

The HapMap Project (along with a previous genome-wide assessment of LD⁷⁷) is a natural extension of the Human Genome Project. Where the reference sequence constructed by the Human Genome Project is informative about the vast majority of bases that are invariant across individuals, the HapMap focuses on DNA sequence differences among individuals. Our understanding of SNP variation and LD around common variants in the sampled populations is reasonably complete; the current picture is unlikely to change with additional data. In other aspects—such as the fine details of local correlation among SNPs, rarer alleles, structural variants, and inter-population differences—these resources are only a first step on the path towards a complete characterization of genetic variation of the human population. Planned extensions of the Phase I map include Phase II of HapMap, with genotyping of another 4.6 million SNPs attempted in the HapMap samples, and detailed genotyping of the HapMap ENCODE regions in additional members of each HapMap population sampled, as well as in samples from additional populations. These results should guide understanding of the robustness and transferability of LD inferences and tag SNPs selected from the current set of HapMap samples.

An important application of the HapMap data is to help make possible comprehensive, genome-wide association studies. There are now laboratory tools that make it practical to undertake such studies, and initial results are encouraging¹³. Given the low prior probability of causality for each SNP in the genome, however, rigorous standards of statistical significance will be needed to avoid a flood of false-positive results. Multiple replications in large samples provide the most straightforward path to identifying robust and broadly relevant associations. Given the potential for confusion if associations of uncertain validity are widely reported (and a persistent tendency towards genetic determinism in public discourse), we urge conservatism and restraint in the public dissemination and interpretation of such studies, especially if non-medical phenotypes are explored. It is time to create mechanisms by which all results of association studies, positive and negative, are reported and discussed without bias.

The success of the HapMap will be measured in terms of the genetic discoveries enabled, and improved knowledge of disease aetiology. Specifically, identifying which genes and pathways are causal in humans has the potential to provide a new and solid foundation for biomedical research. This is equally true whether the variants that lead to the discovery of those genes are themselves rare or common, or of large or small effect. The impact on diagnostics and targeted prevention, however, will depend on how predictive each given allele may be. Where genetic mechanisms underlie treatment

responses, both more efficient trials and individualized preventive and treatment strategies may become practical⁹⁵.

Success identifying alleles conferring susceptibility or resistance to common diseases will also provide a deeper understanding of the architecture of disease: how many genes are involved in each case, whether and how alleles interact with one another⁹⁶ and with environmental exposures to shape clinical phenotypes. In this regard, it will be important to invest heavily in the discovery and characterization of relevant lifestyle factors, environmental exposures, detailed characterization of clinical phenotypes, and the ability to obtain such information in longitudinal studies of adequate size. Where environmental and behavioural factors vary across studies, replication will be hard to come by (as will clinical utility) unless we can learn to capture these variables with the same precision and completeness as genotypic variation. Technological innovation and international collaboration in these realms will probably be required (as they have been in the Human Genome Project and the HapMap) to advance the shared goal of understanding, and ultimately preventing, common human diseases.

METHODS

The project was undertaken by investigators from Japan, the United Kingdom, Canada, China, Nigeria and the United States, and from multiple disciplines: sample collection, sequencing and genotyping, bioinformatics, population genetics, statistics, and the ethical, legal, and social implications of genetic research. The Supplementary Information contains information about project participants and organization.

Choice of DNA samples. Any choice of DNA samples represents a compromise: a single population offers simplicity, but cannot be representative, whereas grid-sampling is representative of the current worldwide population, but is neither practical nor captures historical genetic diversity. The project chose to include DNA samples based on well-known patterns of allele frequencies across populations⁴¹, reflecting historical genetic diversity^{31,32}.

For practical reasons, the project focused on SNPs present at a minor allele frequency (MAF) ≥ 0.05 in each analysis panel, and thus studied a sufficient number of individuals to provide good power for this frequency range³¹. Cell lines and DNA are available at the Coriell Institute for Medical Research (<http://locus.umdnj.edu/nigms/products/hapmap.html>).

Community engagement was employed to explain the project, and to learn how the project was viewed, in the communities where samples were collected^{31,32}. Papers describing the community engagement processes are being prepared.

One JPT sample was replaced for technical reasons, but not in time for inclusion in this report. We surveyed cryptic relatedness among the study participants, and identified a small number of pairs with unexpectedly high allele sharing (Supplementary Information). As the total level of sharing is not great, and as a subset of analyses performed without these individuals were unchanged, we include these individuals in the data and analyses presented here.

Genome-wide SNP discovery. The project required a dense map of SNPs, ideally containing information about validation and frequency of each candidate SNP. When the project started, the public SNP database (dbSNP) contained 2.6 million candidate SNPs, few of which were annotated with the required information.

To generate more SNPs and obtain validation information, shotgun sequencing of DNA from whole-genome libraries and flow-sorted chromosomes was performed³¹, augmented by analysis of sequence traces produced by Applied Biosystems^{97,98}, and information on 1.6 million SNPs genotyped by Perlegen Sciences⁷⁷, including 425,000 not in dbSNP when released (Supplementary Table 7). The HapMap Project contributed about 6 million new SNPs to dbSNP.

At the time of writing (October 2005) dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>) contains 9.2 million candidate human SNPs, of which 3.6 million have been validated by both alleles having been seen two or more times during discovery ('double-hit' SNPs), and 2.4 million have genotype data (Fig. 1).

Comprehensive study of common variation across 5 Mb of DNA. To study patterns of genetic variation as comprehensively as possible, we selected ten 500-kb regions from the ENCODE Project³³. These ten regions were chosen in aggregate to approximate the genome-wide average for G+C content, recombination rate, percentage of sequence conserved relative to mouse sequence, and gene density (Table 2).

In each such region additional sequencing and genotyping were performed to obtain a much more complete inventory of common variation. Specifically, bidirectional PCR-based sequencing was performed across each 500-kb region in

48 individuals (16 YRI, 16 CEU, 8 CHB, 8 JPT). Although the intent was for these same DNA samples to be included in Phase I, eight Yoruba and one Han Chinese sample used in sequencing were not among the 269 samples genotyped. (The nine samples are available from Coriell.)

All variants found by sequencing, and any others in dbSNP (build 121) not found by sequencing, were genotyped in all 269 HapMap samples. If the first attempt at genotyping was unsuccessful, a second platform was tried for each SNP. **False-positive and false-negative rates in PCR-based SNP discovery.** The false-positive rate of SNP discovery by PCR-based resequencing was estimated at 7–11% (for the two sequencing centres), based on genotyping of each candidate SNP in the same samples used for discovery.

The false-negative rate of SNP discovery by PCR resequencing was estimated at 6%, using as the denominator a set of SNPs previously in dbSNP and confirmed by genotyping in the specific individuals sequenced. The false-negative rate was considerably higher, however, for singletons (SNPs seen only as a single heterozygote): 15% of singletons covered by high-quality sequence data were not detected by the trace analysis, and another 25% were missed due to a failure to obtain a high-quality sequence over the relevant base in the one heterozygous individual (D. J. Richter *et al.*, personal communication).

False-positive and false-negative rates in dbSNP. The false positive rate (candidate SNPs that cannot be confirmed as variable sites) estimated for dbSNP was 17%. This represents an upper bound, because dbSNP entries that are monomorphic in the 269 HapMap samples could be rare variants, or polymorphic in other samples. We note that as the catalogue of dbSNP gets deeper, the rate at which candidate SNPs are monomorphic in any given sample is observed to rise (Supplementary Table 8). This is expected because the number of rare SNPs and false positives scales with depth of sequencing, whereas the number of true common variants will plateau.

SNP genotyping for the genome-wide map. Genotyping assays were designed from dbSNP, with priority given to SNPs validated by previous genotyping data or both alleles having been seen more than once in discovery. Data from the Chimpanzee Genome Sequencing Project⁹⁹ were used in SNP validation if they confirmed the ancestral status of a human allele seen only once in discovery (Supplementary Information). Non-synonymous coding SNPs were also prioritized for genotyping. Two whole-genome, array-based genotyping reagents were used efficiently to increase SNP density: 40,000 SNPs from Illumina, and 120,000 SNPs from Affymetrix¹⁰⁰.

To monitor progress, the genome was partitioned into 5-kb bins, with genotyping continuing through iterative rounds until a set of predetermined 'stopping rules' was satisfied in each analysis panel. (1) Minor allele frequency: in each analysis panel a common SNP (MAF ≥ 0.05) was obtained in each 5-kb bin. (2) Spacing: the distance between adjacent SNPs was 2–8 kb, with at least 9 SNPs across 50 kb. (3) 'HapMappable' genome: with available technologies it is challenging to study centromeres, telomeres, gaps in genome sequence, and segmental duplications. The project identified such regions¹⁰¹ (Supplementary Table 9), spanning 4.4% of the finished human genome sequence, in which only a single attempt to develop a genotyping assay was required. (4) Three strikes, you're out: if the above rules were not satisfied after three attempts to develop an assay in a given 5-kb region, or if all available SNPs in dbSNP had been tried, genotyping was considered complete for Phase I. Two attempts were considered sufficient if one attempt was of a SNP previously shown to have MAF ≥ 0.05 in the appropriate population sample in a previous genome-wide survey⁷⁷. (5) Quality control: ongoing and standardized quality control (QC) filters and three rounds of quality assessment (QA) were used to ensure and document the high quality of the genotype data.

QC filters were systematically performed, with each SNP tested for completeness ($>80\%$), consistency across five duplicate genotypes (≤ 1 discrepancy), mendelian inheritance in 60 trios (≤ 1 discrepancy in each of YRI and CEU), and Hardy-Weinberg equilibrium ($P > 0.001^{102}$). SNPs in the Phase I data set passed all the QC filters in all the analysis panels and were polymorphic in the HapMap samples. Failing SNPs were released (with a special flag), as they can help to identify polymorphisms under primers, insertions/deletions, paralogous loci and natural selection.

Three QA exercises were carried out. First, a calibration exercise to 'benchmark' each platform and laboratory protocol. Second, a mid-project evaluation of each genotyping centre. Third, a blind analysis of a random sample of the complete Phase I data set. A number of SNPs were genotyped more than once during the project, or by other investigators, providing additional information about data quality. See the Supplementary Information for full information about the QA exercises.

An exhaustive approach was taken to mtDNA. Alignment of more than 1,000 publicly available mtDNA sequences of African ($n = 87$), European ($n = 928$) and Asian ($n = 238$) geographical origin³⁴ was used to identify 210 common

variants ($MAF \geq 0.05$ in at least one continental region) that were attempted in the samples.

Data release. Data deposited at the Data Coordination Center and released at <http://www.hapmap.org>, a Japanese mirror site <http://hapmap.jst.go.jp/> and dbSNP include ascertainment status of each SNP at the time of selection, primer sequences, protocols for genotyping, genotypes for each sample, allele frequencies, and, for SNPs that failed QC filters, a code indicating the mode(s) of failure.

Initially, because of concern that third parties might seek patents on HapMap data, users were required to agree to a web-based 'click-wrap license', assenting that they would not prevent others from using the data (<http://www.hapmap.org/cgi-perl/registration>). In December 2004 this license was dropped, and all data were released without restriction into the public domain.

Received 11 August; accepted 12 September 2005.

- Lechler, R. & Warrens, A. *HLA in Health and Disease* 2nd edn (Academic Press, San Diego, California, 2005).
- Strittmatter, W. J. & Roses, A. D. Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* **19**, 53–77 (1996).
- Dahlbäck, B. Resistance to activated protein C caused by the factor V R^{506Q} mutation is a common risk factor for venous thrombosis. *Thromb. Haemost.* **78**, 483–488 (1997).
- Altshuler, D. et al. The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
- Deeb, S. S. et al. A Pro12Ala substitution in PPAR γ 2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genet.* **20**, 284–287 (1998).
- Florez, J. C., Hirschhorn, J. & Altshuler, D. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu. Rev. Genomics Hum. Genet.* **4**, 257–291 (2003).
- Begovich, A. B. et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).
- Bottini, N. et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature Genet.* **36**, 337–338 (2004).
- Bell, G. I., Horita, S. & Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**, 176–183 (1984).
- Ueda, H. et al. Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
- Ogura, Y. et al. A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
- Hugot, J. P. et al. Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
- Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
- Haines, J. L. et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
- Edwards, A. O. et al. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
- Puffenberger, E. G. et al. A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease. *Cell* **79**, 1257–1266 (1994).
- Emison, E. S. et al. A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).
- Kerem, B. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
- Hästbacka, J. et al. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
- Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
- Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
- Reich, D. E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
- Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
- Johnson, G. C. et al. Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
- Nickerson, D. A. et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
- Zhu, X. et al. Localization of a small genomic region associated with elevated ACE. *Am. J. Hum. Genet.* **67**, 1144–1153 (2000).
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
- Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
- Patil, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* **5**, 467–475 (2004).
- The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
- Herrnstadt, C. et al. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am. J. Hum. Genet.* **70**, 1152–1171 (2002).
- Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nature Rev. Genet.* **4**, 598–612 (2003).
- The Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339–348 (2002).
- Underhill, P. A. et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001).
- Marchini, J. et al. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* (in the press).
- Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
- Wright, S. *Evolution and the Genetics of Populations Volume 2: the Theory of Gene Frequencies* 294–295 (Univ. of Chicago Press, Chicago, 1969).
- Rosenberg, N. A. et al. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
- Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
- Pe'er, I. et al. Reconciling estimates of linkage disequilibrium in the human genome. *Genome Res.* (submitted).
- Lichten, M. & Goldman, A. S. Meiotic recombination hotspots. *Annu. Rev. Genet.* **29**, 423–444 (1995).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- McVean, G. A. et al. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
- Crawford, D. C. et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36**, 700–706 (2004).
- Kong, A. et al. A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
- Winckler, W. et al. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111 (2005).
- Myers, S. R. & Griffiths, R. C. Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**, 375–394 (2003).
- Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
- Phillips, M. S. et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genet.* **33**, 382–387 (2003).
- Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
- Carlson, C. S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
- de Bakker, P. I. W. et al. Efficiency and power in genetic association studies. *Nature Genet.* Advance online publication, 23 October 2005 (doi:10.1038/ng1669).
- Lin, S., Chakravarti, A. & Cutler, D. J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genet.* **36**, 1181–1188 (2004).
- Weale, M. E. et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A*: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73**, 551–565 (2003).
- Stram, D. O. et al. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36 (2003).
- de la Chapelle, A. & Wright, F. A. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl Acad. Sci. USA* **95**, 12416–12423 (1998).
- Mootha, V. K. et al. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA* **100**, 605–610 (2003).
- Engert, J. C. et al. ARSACS, a spastic ataxia common in northeastern Québec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nature Genet.* **24**, 120–125 (2000).
- Richter, A. et al. Location score and haplotype analyses of the locus for

- autosomal recessive spastic ataxia of Charlevoix-Saguenay, in chromosome region 13q11. *Am. J. Hum. Genet.* **64**, 768–775 (1999).
63. Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85**, 9119–9123 (1988).
 64. Smith, M. W. & O'Brien, S. J. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Rev. Genet.* **6**, 623–632 (2005).
 65. Smith, M. W. *et al.* A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74**, 1001–1013 (2004).
 66. Zhu, X. *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nature Genet.* **37**, 177–181 (2005).
 67. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).
 68. Huang, J. *et al.* Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics* **1**, 287–299 (2004).
 69. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
 70. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
 71. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
 72. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
 73. Singleton, A. B. *et al.* α -Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
 74. McCarroll, S. *et al.* Common deletion variants in the human genome. *Nature Genet.* (in the press).
 75. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
 76. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and recombination hotspots in the human genome. *Science* **310**, 321–324 (2005).
 77. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
 78. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
 79. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
 80. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359**, 794–801 (1992).
 81. Fullerton, S. M., Bernardo Carvalho, A. & Clark, A. G. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142 (2001).
 82. Dawson, E. *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
 83. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
 84. Smith, A. V., Thomas, D. J., Munro, H. M. & Abecasis, G. R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15**, 1519–1534 (2005).
 85. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
 86. Trachtenberg, E. *et al.* Advantage of rare HLA supertype in HIV disease progression. *Nature Med.* **9**, 928–935 (2003).
 87. Pehrson, J. R. & Fuji, R. N. Evolutionary conservation of histone macroH2A subtypes and domains. *Nucleic Acids Res.* **26**, 2837–2842 (1998).
 88. Modrich, P. & Lahue, R. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.* **65**, 101–133 (1996).
 89. Nielsen, R. Human genomics: disclosure of variation. *Nature* **434**, 288–289 (2005).
 90. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature Genet.* **30**, 233–237 (2002).
 91. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
 92. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
 93. Dermitzakis, E. T. *et al.* Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582 (2002).
 94. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
 95. Need, A. C., Motulsky, A. G. & Goldstein, D. B. Priorities and standards in pharmacogenetic research. *Nature Genet.* **37**, 671–681 (2005).
 96. Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703 (2005).
 97. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101**, 1916–1921 (2004).
 98. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
 99. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
 100. Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature Methods* **1**, 109–111 (2004).
 101. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
 102. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
 103. Hill, W. G. & Weir, B. S. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* **54**, 705–714 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: J. Beck, C. Beiswanger, D. Coppock, A. Leach, J. Mintzer and L. Toji (Coriell Institute for Medical Research) for transforming the Yoruba, Japanese and Han Chinese samples, distributing the DNA and cell lines, storing the samples for use in future research, and producing the community newsletters and reports; J. Greenberg and R. Anderson (NIH National Institute of General Medical Sciences) for providing funding and support for cell line transformation and storage in the NIGMS Human Genetic Cell Repository at the Coriell Institute; T. Dibling, T. Ishikura, S. Kanazawa, S. Mizusawa and S. Saito (SNP Research Center, RIKEN) for help with genotyping; C. Hind and A. Moghadam for technical support in genotyping and all members of the subcloning and sequencing teams at the Wellcome Trust Sanger Institute; X. Ke (Wellcome Trust Centre for Human Genetics at the University of Oxford) for help with data analysis; Oxford E-Science Centre for provision of high-performance computing resources; H. Chen, W. Chen, L. Deng, Y. Dong, C. Fu, L. Gao, H. Geng, J. Geng, M. He, H. Li, H. Li, S. Li, X. Li, B. Liu, Z. Liu, F. Lu, F. Lu, G. Lu, C. Luo, X. Wang, Z. Wang, C. Ye and X. Yu (Beijing Genomics Institute) for help with genotyping and sample collection; X. Feng, Y. Li, J. Ren and X. Zhou (Beijing Normal University) for help with sample collection; J. Fan, W. Gu, W. Guan, S. Hu, H. Jiang, R. Lei, Y. Lin, Z. Niu, B. Wang, L. Yang, W. Yang, Y. Wang, Z. Wang, S. Xu, W. Yan, H. Yang, W. Yuan, C. Zhang, J. Zhang, K. Zhang and G. Zhao (Chinese National Human Genome Center at Shanghai) for help with genotyping; P. Fong, C. Lai, C. Lau, T. Leung, L. Luk and W. Tong (University of Hong Kong, Genome Research Centre) for help with genotyping; C. Pang (Chinese University of Hong Kong) for help with genotyping; K. Ding, B. Qiang, J. Zhang, X. Zhang and K. Zhou (Chinese National Human Genome Center at Beijing) for help with genotyping; Q. Fu, S. Ghose, X. Lu, D. Nelson, A. Perez, S. Poole, R. Vega and H. Yonath (Baylor College of Medicine); C. Bruckner, T. Brundage, S. Chow, O. Iartchouk, M. Jain, M. Moorhead and K. Tran (ParAllele Bioscience Inc.); N. Addelman, J. Atilano, T. Chan, C. Chu, C. Ha, T. Nguyen, M. Minton and A. Phong (UCSF) for help with genotyping, and D. Lind (UCSF) for help with quality control and experimental design; R. Donaldson and S. Duan (Washington University) for help with genotyping, and J. Rice and N. Saccone (Washington University) for help with experimental design; J. Wigginton (University of Michigan) for help with implementing and testing QA/QC software; A. Clark, B. Keats, R. Myers, D. Nickerson and A. Williamson for providing advice to NIH; J. Melone, M. Weiss and E. DeHaut-Combs (NHGRI) for help with project management; M. Gray for organizing phone calls and meetings; D. Leja for help with figures; the Yoruba people of Ibadan, Nigeria, the people of Tokyo, Japan, and the community at Beijing Normal University, who participated in public consultations and community engagements; the people in these communities who were generous in donating their blood samples; and the people in the Utah CEPH community who allowed the samples they donated earlier to be used for the Project. We also thank A. Clark, E. Lander, C. Langley and R. Lifton for comments on earlier drafts of the manuscript. This work was supported by the Japanese Ministry of Education, Culture, Sports, Science, and Technology, the Wellcome Trust, Nuffield Trust, Wolfson Foundation, UK EPSRC, Genome Canada, Génome Québec, the Chinese Academy of Sciences, the Ministry of Science and Technology of the People's Republic of China, the National Natural Science Foundation of China, the Hong Kong Innovation and Technology Commission, the University Grants Committee of Hong Kong, the SNP Consortium, the US National Institutes of Health (FIC, NCI, NCR, NEI, NHGRI, NIA, NIAAA, NIAID, NIAMS, NIBIB, NIDA, NIDCD, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NINDS, NLM, OD), the W.M. Keck Foundation, and the Delores Dore Eccles Foundation.

Author Contributions David Altshuler, Lisa D. Brooks, Aravinda Chakravarti, Francis S. Collins, Mark J. Daly and Peter Donnelly are members of the writing group responsible for this manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A. (altshuler@molbio.mgh.harvard.edu) or P.D. (donnelly@stats.ox.ac.uk).

1320

Repetitive shuttling of a motor protein on DNA

Sua Myong¹, Ivan Rasnik¹, Chirlmin Joo¹, Timothy M. Lohman³ & Taekjip Ha^{1,2}

Many helicases modulate recombination, an essential process that needs to be tightly controlled. Mutations in some human disease helicases cause increased recombination, genome instability and cancer. To elucidate the potential mode of action of these enzymes, here we developed a single-molecule fluorescence assay that can visualize DNA binding and translocation of *Escherichia coli* Rep, a superfamily 1 DNA helicase homologous to *Saccharomyces cerevisiae* Srs2. Individual Rep monomers were observed to move on single-stranded (ss)DNA in the 3' to 5' direction using ATP hydrolysis. Strikingly, on hitting a blockade, such as duplex DNA or streptavidin, the protein abruptly snapped back close to its initial position, followed by further cycles of translocation and snapback. This repetitive shuttling is likely to be caused by a blockade-induced protein conformational change that enhances DNA affinity for the protein's secondary DNA binding site, thereby resulting in a transient DNA loop. Repetitive shuttling was also observed on ssDNA bounded by a stalled replication fork and an Okazaki fragment analogue, and the presence of Rep delayed formation of a filament of recombination protein RecA on ssDNA. Thus, the binding of a single Rep monomer to a stalled replication fork can lead to repetitive shuttling along the single-stranded region, possibly keeping the DNA clear of toxic recombination intermediates.

The diverse activities of helicases, such as duplex nucleic acid unwinding¹, protein displacement^{2,3}, and branch migration⁴, are powered by a common engine that translocates directionally on nucleic acids^{5–7}. For example, yeast Srs2 helicase^{8,9} and its bacterial homologues¹⁰ disrupt recombination intermediates formed around ssDNA, probably driven by DNA translocation. We investigated the translocation mechanisms of *E. coli* Rep, an Srs2 homologue that functions in replication restart^{11,12} and replication of certain phages¹³.

We engineered single-cysteine mutants of Rep that retain activity *in vivo* and with and without dye labels *in vitro*¹⁴. Although a Rep monomer cannot unwind DNA *in vitro*^{15–17}, it can translocate on ssDNA in the 3' to 5' direction using ATP hydrolysis¹⁷. The single cysteine was labelled with a donor fluorophore (Cy3) with 90% efficiency¹⁴ and the movement of a donor-labelled Rep on an acceptor (Cy5)-labelled DNA was detected by single-molecule fluorescence resonance energy transfer (FRET)^{18–20}. To optimize the FRET signal, the donor was attached to the position 333 ('leading-edge'²¹) for an acceptor at the 5' end of ssDNA, or to the position 43 ('trailing-edge'²¹) for an acceptor at the 3' end. Double-stranded (ds)DNA (18 base pairs, bp) with a 3' (dT)_n tail (*n* = 40, 60 or 80) and a Cy5 attached to the junction was tethered at the duplex end to a polymer-coated quartz slide via biotin–streptavidin, and single-molecule data were obtained in the presence of 300 pM of Rep and 1 mM ATP in solution using dual-view wide-field total-internal-reflection fluorescence microscopy^{14,16,22} with 15-ms time resolution (Fig. 1a and Supplementary Fig. S1). Direct excitation of the acceptor at the donor-excitation wavelength of 532 nm is insignificant and the polymer coating eliminates nonspecific surface binding of proteins^{14,16,23}, so single-molecule fluorescence signals are observed only when the protein binds the DNA.

Blockade-induced repetitive shuttling of a Rep monomer on DNA

When a Rep monomer (Cy3 labelled at position 333) binds the partial duplex DNA with a (dT)₈₀ tail, the donor fluorescence signal

rises abruptly, combined with a weak acceptor signal (Fig. 1b). This is followed by a gradual decrease in donor signal and a concomitant gradual increase in acceptor signal (and corresponding FRET increase, Fig. 1d), consistent with ssDNA translocation in the 3' to 5' direction towards the junction. Since Rep cannot unwind duplex DNA as a monomer *in vitro*^{15,16}, the junction presents itself as a blockade at which the protein is expected to stop and dissociate. Instead of the anticipated dissociation, however, we observed an instantaneous (within 15 ms) FRET decrease to near the initial value (Fig. 1b and d) which is followed by further cycles of a gradual FRET increase and an abrupt FRET decrease. This sawtooth-shaped cycle was repeated several times until it was finally terminated by protein dissociation or photobleaching. We interpret the sawtooth pattern as reflecting repeated cycles of ssDNA translocation followed by the protein snapback to near its initial binding region (see below) and will call it 'repetitive shuttling' henceforth. Repetitive shuttling was observed over a wide range of solution conditions (15–100 mM NaCl, 2.1–10 mM MgCl₂, 22–37 °C) and also with DNA containing ssDNA tails of mixed sequences (Fig. 1c, Supplementary Figs S2, S3 and S5). Typically, more than 80% of binding events resulted in repetitive shuttling (Supplementary Fig. S2).

Several lines of evidence strongly suggest that the sawtooth pattern is caused by a single Rep monomer rather than successive binding of different monomers. (1) Because the labelling efficiency is about 90%, a single monomer can be discerned by the well-defined fluorescence intensity of a single donor¹⁴. (2) Sawtooth patterns are observed as well-isolated bursts (Fig. 1b and c). (3) When a flow of buffer devoid of protein was applied during data acquisition to remove free proteins in solution, the sawtooth patterns persisted.

Figures 1d–f show histograms of the time between two successive snapbacks, Δt . The peak of the histogram shifts to longer times as the tail length increases (0.62, 1.0 and 1.23 s for 40, 60 and 80 nucleotide (nt) tails, respectively), suggesting that the gradual FRET increase corresponds to ssDNA translocation. Single-molecule FRET time

¹Physics Department, University of Illinois, Urbana-Champaign, and ²Howard Hughes Medical Institute, Urbana, Illinois 61801, USA. ³Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, Missouri 63110, USA.

traces show that the protein spends longer times in the low FRET phase for longer ssDNA tails, consistent with this interpretation (Figs 1d–f). A sawtooth pattern with a 2.7 s period was observed using a similar DNA with a 182-nt 3' tail of mixed sequence (Supplementary Fig. S3), further supporting our interpretation. The period of the sawtooth pattern increases at low ATP concentrations or when 10–20 μ M ATP γ S is added to a high concentration of ATP (Supplementary Fig. S4), strongly implicating ssDNA translocation powered by ATP hydrolysis. The closest distance between the two dyes is about 4–5 nm (FRET efficiency $\sim$ 0.75 and R_0 of 5–6.3 nm; refs 14, 23), consistent with the structure of another superfamily 1 DNA helicase, PcrA of *Bacillus stearothermophilus*, bound to a partial duplex junction²⁴. The remarkable regularity of the sawtooth pattern and narrowly peaked Δt histograms suggest that the site for the re-initiation of translocation is not random and is likely to be localized near the 3' end. The linear relation between the period and the tail length further suggests that snapback redirects the protein primarily to a region near the 3' end. The ssDNA translocation rates estimated from the single-molecule experiments agree with those obtained from ensemble studies with unmodified Rep when both are performed under similar conditions (Supplementary Fig. S5).

We observed similar sawtooth patterns (average period = 1.2 s) when a ssDNA, (dT)₅₀, is terminated at the 5' end by a biotin bound to a streptavidin (Fig. 1g and h). Because the acceptor is near the 3' end, we observe a gradual FRET decrease during translocation followed by an abrupt FRET increase. In contrast, only single translocation events followed by dissociation were observed when the 3' end is attached to a streptavidin and the 5' end is free (Fig. 1i and j). These observations suggest that the encounter of a physical blockade such as duplex DNA junction or streptavidin may trigger a snapback.

In contrast to the previously observed backward movements of RecBCD²⁵ and UvrD²⁶, repetitive shuttling of Rep is (1) deterministic, that is, the forward and backward movements are repeated in

regular intervals and are reliably triggered by a blockade, (2) highly asymmetric so that the backward movement occurs too fast to be resolved, and (3) observed without applied force.

Physical mechanisms of repetitive shuttling

Rep binds to ssDNA with a rate constant $\sim$ 20 times slower than the diffusion-limited value¹⁵, so a Rep that dissociates completely from DNA is much more likely to diffuse away than to rebind to the same DNA. Therefore, repetitive shuttling is unlikely to be due to complete dissociation and rebinding of Rep. To remain bound to the DNA during snapback, the protein either has to slide all the way towards the 3' end in less than 15 ms, or has to make simultaneous contacts with the junction and the 3' end. We favour the latter mechanism, on the basis of the following studies of DNA conformations during translocation (Fig. 2a). In the presence of unlabelled Rep and ATP, a DNA with a 3' (dT)₄₀ tail labelled at the near extremities with a donor and an acceptor showed mostly low FRET values ($\sim$ 0.4) but with brief, regular spikes to high FRET efficiency ($\sim$ 0.7; spikes had average duration 0.17 s) (Fig. 2b). No such spikes were observed from DNA alone. The average period of this pattern is 0.7 s, very similar to the period of the sawtooth pattern observed with labelled Rep translocating along a 40-nt tail. We therefore interpret the high FRET spike as reflecting the simultaneous binding of Rep to both the junction and the 3' end of the ssDNA, resulting in the transient formation of a DNA loop. This mechanism requires at least two distinct DNA binding sites on the Rep monomer, the primary binding site as observed in crystal structures^{24,27} and the secondary binding site for the 3' end. We suggest that a blockade encounter induces Rep conformational changes that increase the DNA affinity of the secondary binding site. The loop formation can follow immediately because ssDNA is highly flexible and its conformational fluctuations are much faster than our time resolution²⁸.

We next investigated whether Rep undergoes conformational changes coupled with repetitive shuttling. Rep is structurally homologous to another superfamily 1 helicase PcrA, and is composed of

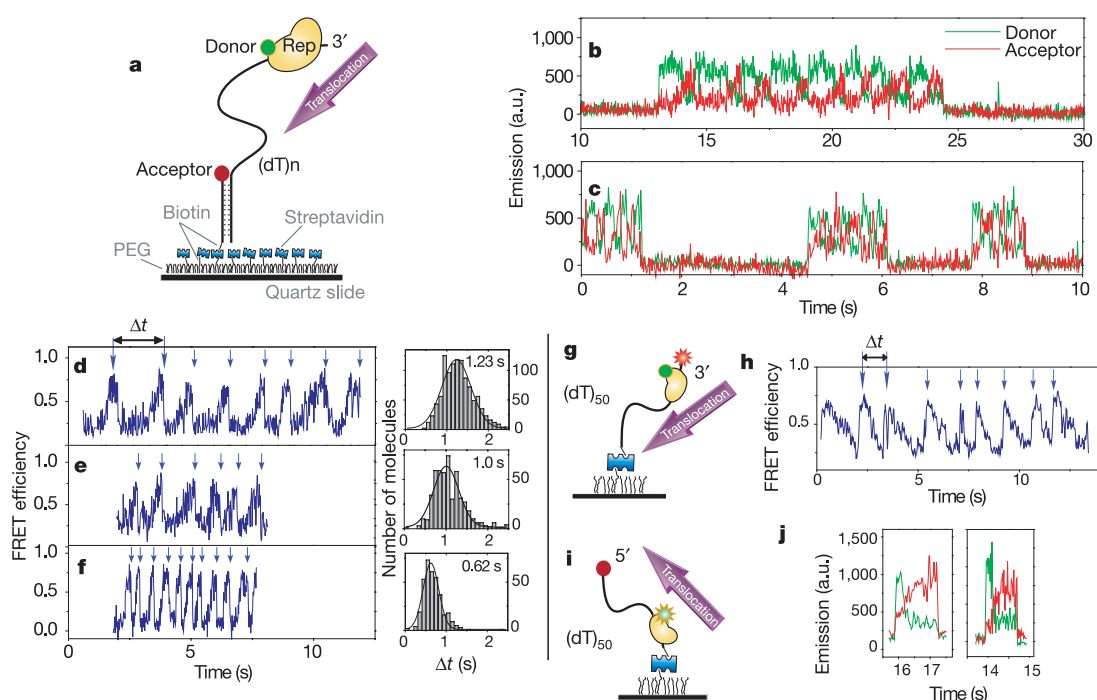


Figure 1 | Blockade-induced repetitive shuttling. **a**, Donor-labelled Rep binds to a 3' ssDNA tail and translocates towards the acceptor. **b**, **c**, Fluorescence intensity traces for a 3' (dT)₈₀ tail at 22 °C (**b**) and 37 °C (**c**). **d–f**, FRET traces for 3' tails of 80, 60 and 40 dTs. Histograms of Δt between snapbacks (arrows) are shown with gaussian fits (solid lines).

g, Donor-labelled Rep moves away from the acceptor towards streptavidin. **h**, Repeated cycles of gradual decrease and abrupt increase of FRET. **i**, Donor-labelled Rep moves towards the acceptor and away from streptavidin. **j**, Only gradual FRET increase was observed. a.u., arbitrary units.

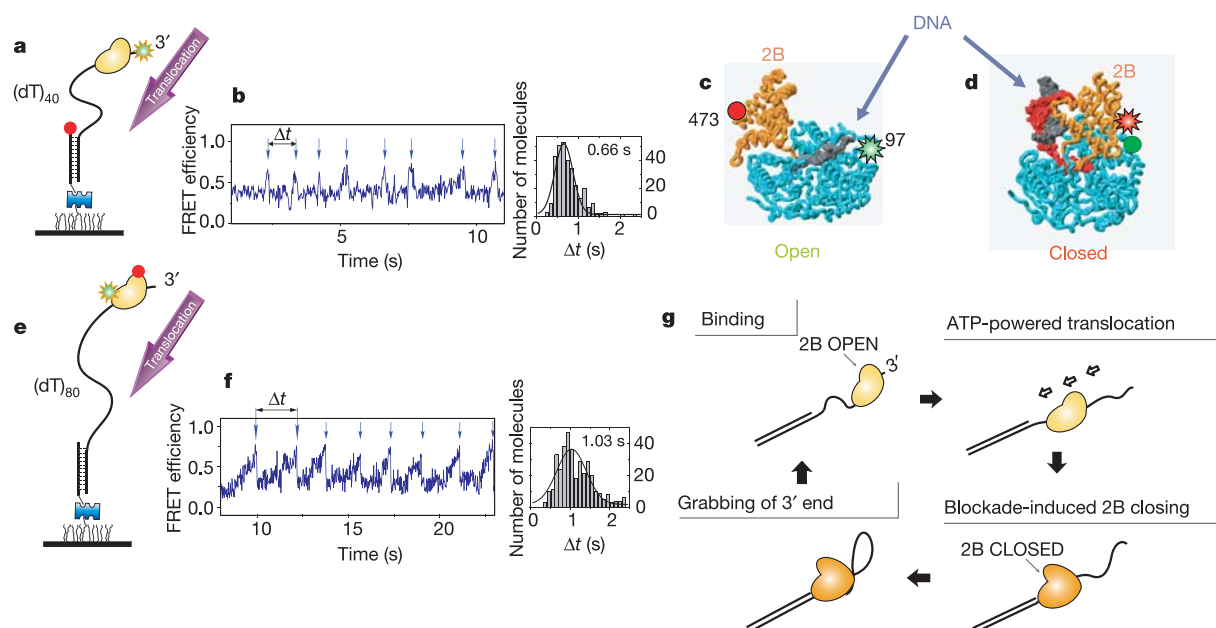


Figure 2 | Physical mechanism of repetitive shuttling. **a**, Unlabelled Rep moves on a dual-labelled $(dT)_{40}$ tail. **b**, FRET trace shows brief spikes to high FRET (arrows); the Δt histogram shows a gaussian fit (solid line). **c**, Crystal structure of Rep (2B in orange and the rest in blue) bound to ssDNA (grey). Red and green symbols denote cysteines at residues 473 and 97. **d**, PcrA structure bound to a partial duplex DNA. Residues equivalent to residues

473 and 97 of Rep are marked with red and green symbols. **e**, Dual-labelled Rep moves on a $(dT)_{80}$ tail. **f**, Repeated cycles of gradual FRET increase and abrupt FRET decrease. The Δt histogram with a gaussian fit (solid line) is shown. **g**, Rep undergoes conformational changes upon blockade approach, transfer to 3' end via a DNA loop, and restart of translocation.

four subdomains (1A, 2A, 1B and 2B)^{24,27}. The 2B subdomain of Rep is dispensable for unwinding²⁹ and ssDNA translocation¹⁷. Rep was crystallized in two forms, open (Fig. 2c) and closed, which differ in the 2B subdomain orientation²⁷. A PcrA bound to a 3'-tailed dsDNA was crystallized in the closed form (Fig. 2d)²⁴ and Rep bound to a 3'-tailed dsDNA in solution favours the closed form¹⁴. To test whether the 2B subdomain closes as Rep approaches the junction, we engineered a double-cysteine mutant of Rep (positions 97 and 473

on the 1B and 2B subdomains, respectively) and labelled it stochastically with Cy3 and Cy5 so that 2B closing would result in a FRET increase (Fig. 2c and d). Single-molecule measurements could identify this mutant labelled with one donor and one acceptor³⁰ as it moves on unlabelled DNA with a 3' $(dT)_{80}$ tail (Fig. 2e). A representative time trace in Fig. 2f shows several cycles of gradual FRET increase and an abrupt FRET decrease. The shortest distance between the donor and acceptor fluorophores on Rep is estimated to

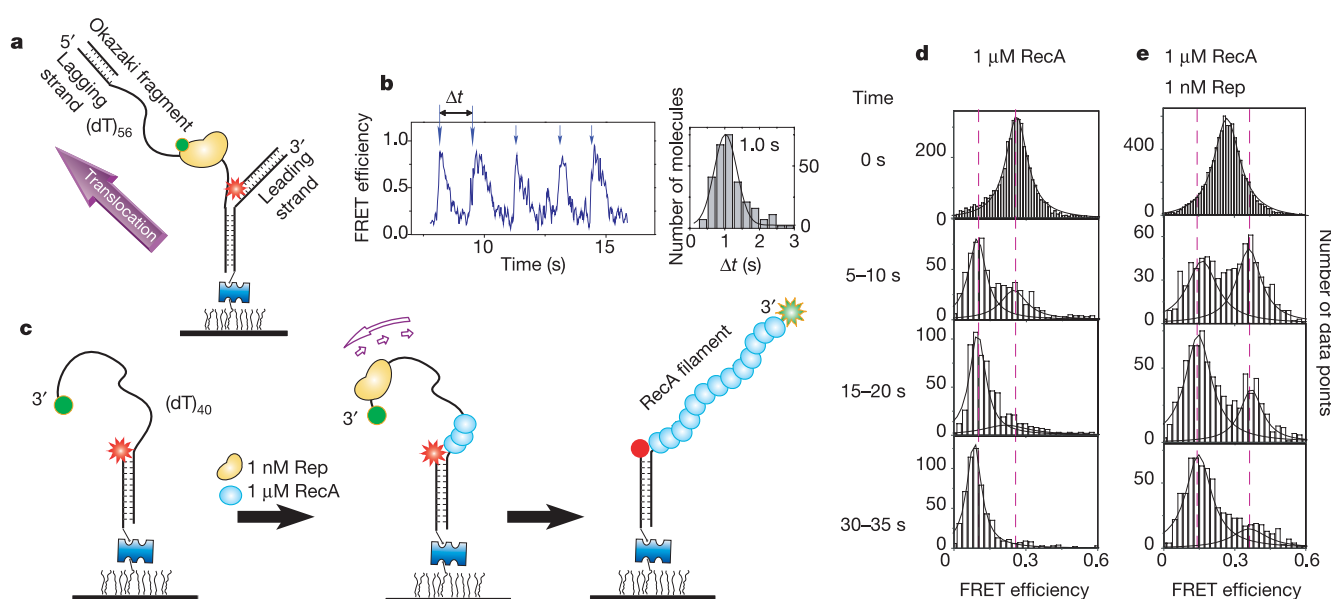


Figure 3 | Potential roles of repetitive shuttling. **a**, Donor-labelled Rep moves on $(dT)_{56}$ between a replication fork and an Okazaki fragment analogue. **b**, FRET trace shows cycles of gradual decrease/abrupt increase. The Δt histogram with a gaussian fit (solid line) is shown. **c**, FRET detection of RecA filament formation on $(dT)_{40}$, which may be hindered by Rep.

d, Time-dependent single-molecule FRET histograms of the DNA shown in **c** after adding RecA and ATP. Dashed lines denote the FRET values for the DNA only and the RecA filament. Also shown are the lorentzian fits. **e**, Same as in **d**, except for the inclusion of 1 nM Rep. The shift to higher FRET is probably due to Rep activity.

be 4–5 nm (FRET value ~ 0.75), in agreement with the closed form of the Rep structure (3-nm C_α distance plus dye linkers). Similar cycles of FRET changes were observed when the doubly labelled Rep mutant was moving on a (dT)₆₀ attached to a streptavidin via 5'-biotin (Supplementary Fig. S6). Thus, the 2B subdomain closes gradually as the protein approaches a blockade, and its complete closing may correlate with the affinity enhancement of the secondary binding site towards the 3' end of the ssDNA, followed by the snapback and the restart of translocation (Fig. 2g and Supplementary Fig. S1). Single-molecule measurements with various ATP analogues suggest that the 2B subdomain opens and closes during each ATP hydrolysis cycle (our unpublished observations). Although the gradual FRET change observed here is surprising, it may reflect a rapid equilibration between the two conformations of the 2B domain whose midpoint shifts as the blockade is approached.

Potential roles of repetitive shuttling in replication restart

Next, we designed a DNA substrate that is relevant to Rep's function in replication restart (Fig. 3a)^{11,12}. Rep binds with high affinity to a three-way junction with a 5' overhang¹² that resembles a stalled replication fork with an incompletely synthesized lagging strand, suggesting that Rep may recognize a fork ready to be restarted. However, the role of Rep as a helicase is not clear in this context because its 3'–5' translocation/unwinding activity would result in translocation towards the Okazaki fragment rather than unwinding the duplexes at the fork. The inability to unwind DNA (the Okazaki fragment) as a monomer, the high affinity for the fork structure, and the ability to snap back quickly may combine to allow a Rep monomer to shuttle back and forth multiple times on the ssDNA region before dissociation. Indeed, we observed the sawtooth pattern from such a structure with a ssDNA gap, (dT)₅₆, and an Okazaki fragment analogue (16 bp) (Fig. 3a and b). The time trace in Fig. 3b shows clear evidence of repetitive shuttling (period 1.0 s). This also shows that a free 3' end is not the only DNA structure on which a snapback can be observed. Repetitive shuttling was also observed when the ssDNA gap was 56 nt of mixed sequence (Supplementary Fig. S7).

What might be the biological role of repetitive shuttling of Rep? Rep functions in the restart of stalled replication forks^{11,12} and in the replication of certain phages³¹, but the *in vivo* functional form (monomer, dimer, and so on) of this low-copy-number³² protein is not known. Two other superfamily 1 helicases, yeast Srs2 (refs 8, 9) and *E. coli* UvrD¹⁰, can displace Rad51 and RecA presynaptic filaments from ssDNA, respectively. Deletion of both Rep and UvrD is lethal in *E. coli*, and it was suggested that Rep prevents the formation of potentially toxic RecA filaments and that UvrD destroys the filament in the absence of Rep¹⁰. Such a preventative role of Rep is suggested by Rep's ability to interfere with RecA filament formation (Fig. 3d and e). Filament formation by RecA (at [RecA] = 1 μ M) on (dT)₄₀ was monitored by FRET (Fig. 3c) and was observed to be delayed substantially even at Rep concentrations as low as 1 nM. *E. coli* remains viable even with the deletion of both Rep and UvrD if RecFOR machinery is defective³³. RecFOR removes SSB from stalled replication forks and loads RecA³⁴, so the lethality of the double deletion of Rep and UvrD may indeed arise from uncontrolled recombination via the RecA filaments. On the basis of the present results, we propose that repetitive shuttling of Rep may be an effective means of keeping the ssDNA clear of unwanted proteins. Such a mode of action does not require the canonical helicase function of duplex unwinding, and therefore could be carried out by a Rep monomer. Whether a Rep monomer can indeed perform these functions *in vivo* is yet to be determined.

METHODS

Proteins were purified and labelled as described¹⁴. A quartz slide was coated with poly-ethylene glycol (PEG) and streptavidin as described^{14,16,23}. After immobilizing biotinylated DNA (300 pM), images were obtained in the presence of 300 pM

Rep in solution in a wide-field total-internal-reflection fluorescence microscope^{14,16,23} with 15-ms time resolution using an electron multiplying charge-coupled device (CCD) camera (iXon DV 887-BI, Andor Technology) and a homemade C++ program written by S. A. McKinney (available on request). The Rep concentration used is much lower than the ~ 300 nM needed for efficient dimer formation and DNA unwinding *in vitro*¹⁵. All measurements were performed at 22 °C with the following buffer composition unless mentioned otherwise: 10 mM Tris-HCl, pH 7.6, 1 mM ATP, 12 mM MgCl₂, 15 mM NaCl, 10% glycerol (v/v), and an oxygen scavenger system¹⁴ to slow photobleaching. Snapback events were visually identified and their timings were recorded using a MATLAB program (available on request). Single-molecule FRET histograms for RecA experiments were obtained by averaging over 1 s. RecA concentration used is within the range expected *in vivo*. FRET values were calculated as the ratio between the acceptor intensity and the total intensity. DNA sequences, modifications and annealing procedures are described in the Supplementary Information.

Received 15 May; accepted 21 July 2005.

- Lohman, T. M. & Bjornson, K. P. Mechanisms of helicase-catalyzed DNA unwinding. *Annu. Rev. Biochem.* **65**, 169–214 (1996).
- Byrd, A. K. & Raney, K. D. Protein displacement by an assembly of helicase molecules aligned along single-stranded DNA. *Nature Struct. Mol. Biol.* **11**, 531–538 (2004).
- Fairman, M. E. *et al.* Protein displacement by DEXH/D "RNA helicases" without duplex unwinding. *Science* **304**, 730–734 (2004).
- Kaplan, D. L. & O'Donnell, M. DnaB drives DNA branch migration and dislodges proteins while encircling two DNA strands. *Mol. Cell* **10**, 647–657 (2002).
- Lee, M. S. & Mariani, K. J. Differential ATP requirements distinguish the DNA translocation and DNA unwinding activities of the *Escherichia coli* PRI A protein. *J. Biol. Chem.* **265**, 17078–17083 (1990).
- Kawaoka, J., Jankowsky, E. & Pyle, A. M. Backbone tracking by the SF2 helicase NPH-II. *Nature Struct. Mol. Biol.* **11**, 526–530 (2004).
- von Hippel, P. H. Helicases become mechanistically simpler and functionally more complex. *Nature Struct. Mol. Biol.* **11**, 494–496 (2004).
- Krejci, L. *et al.* DNA helicase Srs2 disrupts the Rad51 presynaptic filament. *Nature* **423**, 305–309 (2003).
- Veaute, X. *et al.* The Srs2 helicase prevents recombination by disrupting Rad51 nucleoprotein filaments. *Nature* **423**, 309–312 (2003).
- Veaute, X. *et al.* UvrD helicase, unlike Rep helicase, dismantles RecA nucleoprotein filaments in *Escherichia coli*. *EMBO J.* **24**, 180–189 (2005).
- Sandler, S. J. Multiple genetic pathways for restarting DNA replication forks in *Escherichia coli* K-12. *Genetics* **155**, 487–497 (2000).
- Mariani, K. J. Mechanisms of replication fork restart in *Escherichia coli*. *Phil. Trans. R. Soc. Lond. B* **359**, 71–77 (2004).
- Scott, J. F., Eisenberg, S., Bertsch, L. L. & Kornberg, A. A mechanism of duplex DNA replication revealed by enzymatic studies of phage ϕ X174: catalytic strand separation in advance of replication. *Proc. Natl Acad. Sci. USA* **74**, 193–197 (1977).
- Rasnik, I., Myong, S., Cheng, W., Lohman, T. M. & Ha, T. DNA-binding orientation and domain conformation of the *E. coli* Rep helicase monomer bound to a partial duplex junction: Single-molecule studies of fluorescently labelled enzymes. *J. Mol. Biol.* **336**, 395–408 (2004).
- Cheng, W., Hsieh, J., Brendza, K. M. & Lohman, T. M. *E. coli* Rep oligomers are required to initiate DNA unwinding *in vitro*. *J. Mol. Biol.* **310**, 327–350 (2001).
- Ha, T. *et al.* Initiation and reinitiation of DNA unwinding by the *Escherichia coli* Rep helicase. *Nature* **419**, 638–641 (2002).
- Brendza, K. M. *et al.* Auto-inhibition of *E. coli* Rep monomer helicase activity by its 2B sub-domain. *Proc. Natl Acad. Sci. USA* **102**, 10081 (2005).
- Ha, T. *et al.* Probing the interaction between two single molecules—fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* **93**, 6264–6268 (1996).
- Weiss, S. Fluorescence spectroscopy of single biomolecules. *Science* **283**, 1676–1683 (1999).
- Ha, T. Single molecule fluorescence resonance energy transfer. *Methods* **25**, 78–86 (2001).
- Mukhopadhyay, J. *et al.* Translocation of σ^{70} with RNA polymerase during transcription: fluorescence resonance energy transfer assay for movement relative to DNA. *Cell* **106**, 453–463 (2001).
- Zhuang, X. W. *et al.* A single-molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051 (2000).
- Blanchard, S. C., Kim, H. D., Gonzalez, R. L. Jr, Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101**, 12893–12898 (2004).
- Velankar, S. S., Soultanas, P., Dillingham, M. S., Subramanya, H. S. & Wigley, D. B. Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* **97**, 75–84 (1999).
- Perkins, T. T., Li, H. W., Dalal, R. V., Gelles, J. & Block, S. M. Forward and reverse motion of single RecBCD molecules on DNA. *Biophys. J.* **86**, 1640–1648 (2004).

26. Dessinges, M. N., Lionnet, T., Xi, X. G., Bensimon, D. & Croquette, V. Single-molecule assay reveals strand switching and enhanced processivity of UvrD. *Proc. Natl Acad. Sci. USA* **101**, 6439–6444 (2004).
27. Korolev, S., Hsieh, J., Gauss, G. H., Lohman, T. M. & Waksman, G. Major domain swiveling revealed by the crystal structures of complexes of *E. coli* Rep helicase bound to single-stranded DNA and ATP. *Cell* **90**, 635–647 (1997).
28. Murphy, M. C., Rasnik, I., Cheng, W., Lohman, T. M. & Ha, T. Probing single stranded DNA conformational flexibility using fluorescence spectroscopy. *Biophys. J.* **86**, 2530–2537 (2004).
29. Cheng, W. *et al.* The 2B domain of the *Escherichia coli* Rep protein is not required for DNA helicase activity. *Proc. Natl Acad. Sci. USA* **99**, 16006–16011 (2002).
30. Margittai, M. *et al.* Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proc. Natl Acad. Sci. USA* **100**, 15516–15521 (2003).
31. Denhardt, D. T., Dressler, D. H. & Hathaway, A. The abortive replication of ϕ X174 DNA in a recombination deficient mutant of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **57**, 813–820 (1967).
32. Scott, J. F. & Kornberg, A. Purification of the rep protein of *Escherichia coli*. An ATPase which separates duplex DNA strands in advance of replication. *J. Biol. Chem.* **253**, 3292–3297 (1978).
33. Petit, M. A. & Ehrlich, D. Essential bacterial helicases that counteract the toxicity of recombination proteins. *EMBO J.* **21**, 3137–3147 (2002).
34. Morimatsu, K. & Kowalczykowski, S. C. RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. *Mol. Cell* **11**, 1337–1347 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. A. McKinney for writing the data acquisition program and the National Institute of Health for grants (to T.H. and T.M.L.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.H. (tjha@uiuc.edu).

LETTERS

How Prometheus creates structure in Saturn's F ring

Carl D. Murray¹, Carlos Chavez¹, Kevin Beurle¹, Nick Cooper¹, Michael W. Evans¹, Joseph A. Burns² & Carolyn C. Porco³

Images of Saturn's narrow and contorted F ring returned by the Cassini spacecraft¹ have revealed phenomena not previously detected in any planetary ring system. The perturbing effect of the inner shepherding satellite, Prometheus, seems to introduce channels through the F ring and a 'streamer'—a line of particles that link the ring to the satellite. The detailed mechanism for the formation of these features has been lacking an explanation. Here we show that these phenomena can be understood in terms of a simple gravitational interaction as Prometheus approaches and recedes from the F ring every 14.7 hours. Our numerical models show that as Prometheus recedes from its closest approach to the F ring, it draws out ring material; one orbital period later, this affected region has undergone keplerian shear and is visible as a

channel, in excellent agreement with structures seen in the Cassini images. Prometheus' periodic disruption of the F ring will become more pronounced as the two orbits approach their minimum separation in 2009. The model predicts that the appearance of streamers and the associated channels will vary in a regular fashion on a timescale of one orbital period.

As the closer and more massive of the F ring's two shepherding satellites, Prometheus was always the best candidate for being the narrow ring's major perturber^{2–4}. At the time of Cassini's Saturn Orbit Insertion (SOI) on 1 July 2004, the minimum distance between Prometheus⁵ and the F ring's core⁶ was ~500 km. However, the SOI images¹ (Fig. 1a) obtained by the Cassini Imaging Science Subsystem⁷ (ISS) confirmed Voyager observations⁸ of an F ring with at least two

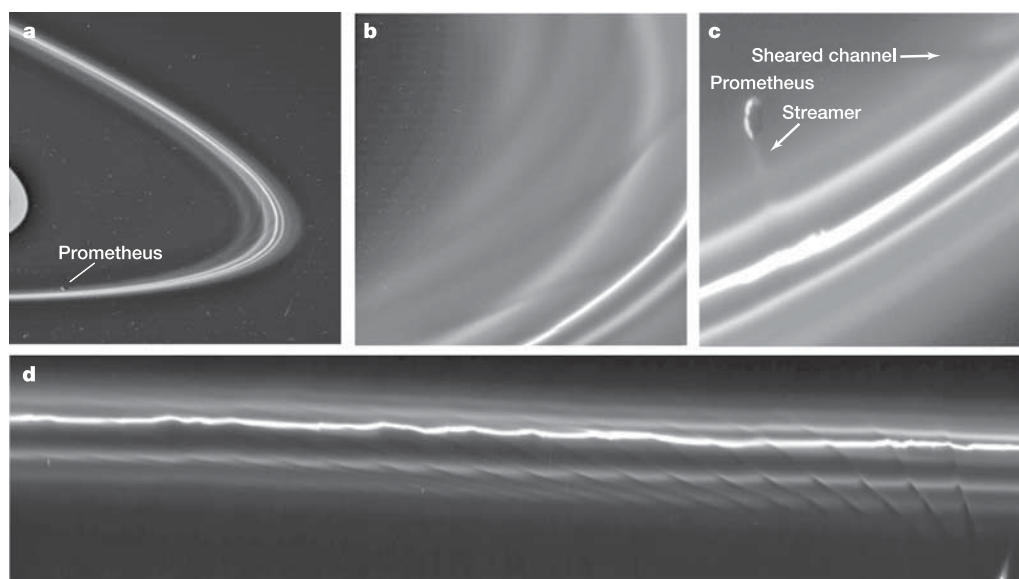


Figure 1 | Cassini images of Saturn's F ring. **a**, Part of an image (W1467350440) of Saturn's F ring taken on 1 July 2004 by the Cassini ISS wide angle camera. The position of Prometheus is indicated. The ring has multiple strands, and the ansa region (at the right) shows 'drape' structure in material that has recently passed Prometheus. **b**, An image (N1467350440) taken by the Cassini ISS narrow angle camera at the same time as **a**, showing the inner part of the F ring near the ansa region. The material interior to the F ring's core (the narrow, bright strand) shows detailed, regular structure with multiple channels giving rise to the 'drape' appearance. **c**, Close-up of a Cassini ISS narrow angle camera image (N1477737741) taken on 29 October 2004 showing Prometheus, the F ring, a streamer of material appearing to

connect them and a sheared channel. **d**, A mosaic of 15 reprojected images of the F-ring region taken by the Cassini ISS narrow angle camera on 13 April 2005. The reprojection procedure begins with a raw digital image that has been pointed using background stars. With the assumption that the ring material is in the equatorial plane, the centre of each pixel corresponds to a unique radius and longitude in a standard reference frame. The image covers ~60° in longitude and the radial range is 1,500 km. The slope of the ring strands is a consequence of the eccentricity of the ring becoming evident over the longitude range. The period of time between the first and last images in the mosaic is ~2.5 h. The mosaic shows multiple sheared channels to the left of Prometheus.

¹Astronomy Unit, Queen Mary, University of London, Mile End Road, London E1 4NS, UK. ²Department of Astronomy & Department of Theoretical and Applied Mechanics, Cornell University, Space Sciences Building, Ithaca, New York 14853, USA. ³Cassini Imaging Central Laboratory for Operations, Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, Colorado 80301, USA.

additional strands and a background dust population that extends across ≈ 700 km in radius. Therefore Prometheus, while not reaching the F ring's core, penetrates the inner dust region each orbital period of Prometheus (14.7 h). A radius versus longitude reprojection of an ISS image taken by the narrow angle camera suggests that the appearance of 'drape' structure¹ (Fig. 1b) resulted from a regular sequence of gap openings or channels in the inner dust sheet, each separated by the spacings of Prometheus' closest approach. If a radial

feature were produced at the time and longitude of previous apoapse passages of Prometheus (when the satellite was at its maximum distance from Saturn), then each channel would shear out, producing the sort of gradient that is visible.

Early numerical models of the system^{9–11} showed the complicated structures that could result and gave important insight into the dynamics of such encounters, but the existence of 'streamers' had not been suspected on the basis of either the models or the Voyager

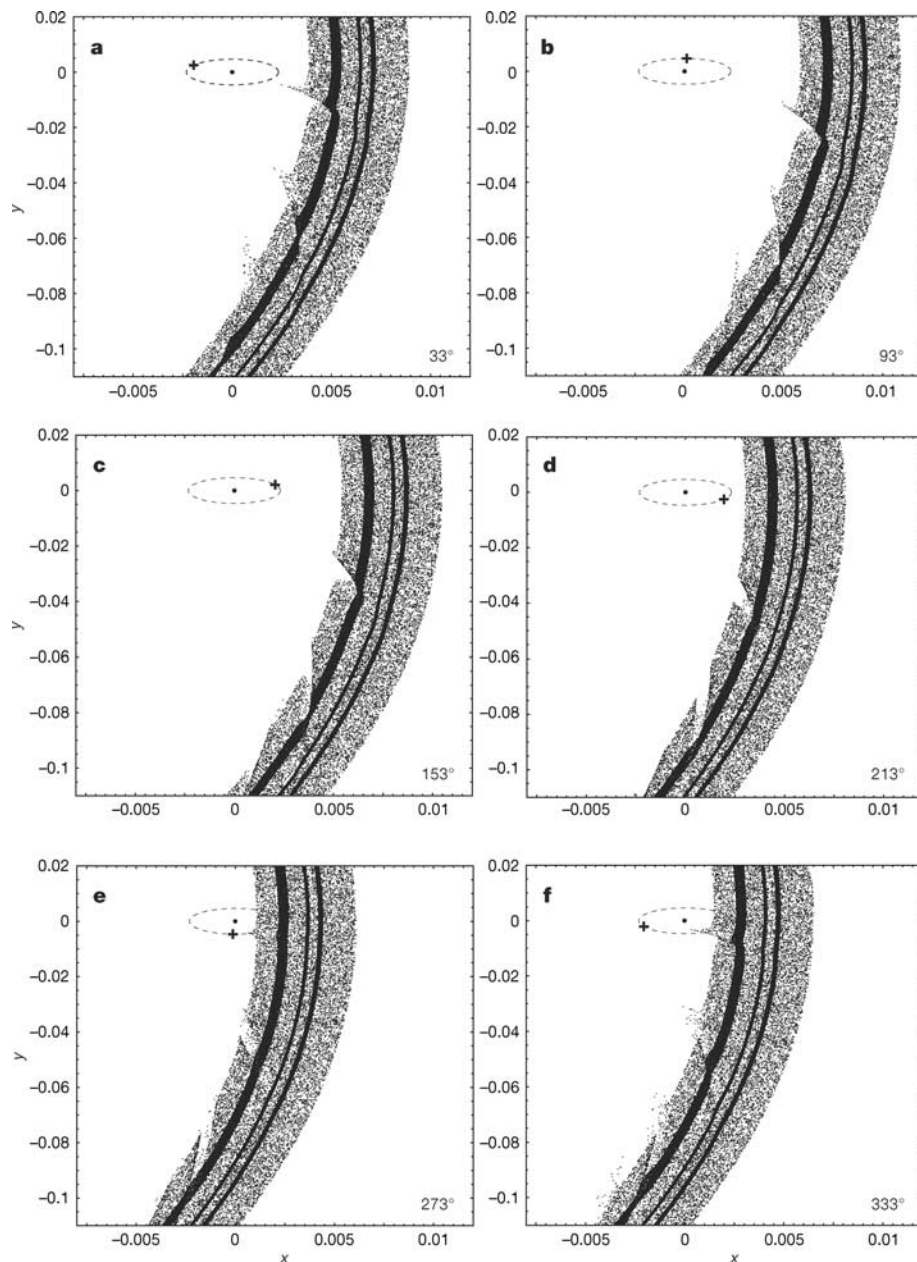


Figure 2 | Numerical modelling of an encounter between Prometheus and the F ring. In each case, the positions of the test particles and Prometheus are plotted in the plane of Prometheus' orbit and in a frame rotating with its mean motion^{1,5} ($587.285^\circ\text{d}^{-1}$). All objects are moving in direct orbits (upward or anticlockwise on the figure), but the use of a rotating frame means that the ring particles (all with semimajor axes larger than that of Prometheus) will move more slowly than it and pass from the top to the bottom of each plot. Both the ring ($e = 0.00254$) and the satellite are in eccentric orbits, with Prometheus' apoapse differing from the F ring's periapse by 66.5° corresponding to the configuration at the chosen epoch. The dashed ellipse is the path of Prometheus in this reference frame and the cross (+) in each plot denotes the position of the centre of Prometheus; x

points radially outwards from Prometheus while +y is in the direction of overall orbital motion. The plots are for different values of the mean anomaly, M , an angle that increases linearly with time such that $M = 0^\circ$ corresponds to periapse and $M = 180^\circ$ to apoapse. The plots correspond to values of Prometheus' M of 33° (a), 93° (b), 153° (c), 213° (d), 273° (e) and 333° (f). The unit of distance is Prometheus' semimajor axis, and the origin (denoted by a filled circle) is chosen to be Prometheus' centre of epicyclic orbital motion¹¹ in the rotating frame. Note that the x and y scales are different; this corresponds to viewing the system from an elevation angle of $\sim 9.5^\circ$ above the ring plane. The different placement of the F ring within each frame is a consequence of the F ring's eccentricity.

images. The realization that Prometheus (orbital semimajor axis, a , 139,700 km; eccentricity, e , 0.0023) could enter parts of the F ring^{2,4}, combined with improved measurements of the F ring's structure⁸, led to a more realistic model¹² that showed streamers even when the long axes of the orbits of Prometheus and the F ring were not anti-aligned. The phenomenon has now been detected in Cassini images¹ (Fig. 1c). We have undertaken a new numerical simulation that makes use of data on the F ring derived from the Cassini images obtained at SOI¹.

The system was modelled with a three-dimensional numerical integration of a spherical Prometheus¹³ (radius 50 km, mass 2.11×10^{17} kg) perturbing 200,000 test particles arranged in four radial groups of 50,000 each. The groupings contained three strands with mean semimajor axes of 140,084 km, 140,224 km and 140,314 km with radial widths of 70 km, 20 km and 30 km, respectively. These values were derived from radial measurements made from the ISS narrow angle images¹ obtained at SOI. In addition, we included a 700-km-wide background sheet of particles with a mean semimajor axis and eccentricity equal to that of the F ring. In each group, the particles were assigned random semimajor axes in a uniform distribution across the group's width and random initial orbital longitudes covering 23° . All other initial orbital elements were taken to be those of the F ring⁴ at the chosen epoch (Julian date 2453187.71, corresponding to UTC 4:56:25 on 1 July 2004). The range of longitudes was chosen in order that Prometheus would encounter the F ring five times in the course of the integration, thereby allowing us to see the effects on the ring of a series of encounters with Prometheus. Because we were interested in the short-term evolution

of the system, we did not include the dynamical effects of Saturn's oblateness. Although the orbits of the strands and Prometheus are all precessing owing to these non-spherical components of the planet's gravity field, their relative precession rate is only $0.057^\circ \text{d}^{-1}$ because of their close proximity. Consequently precessional effects are not important on the ~ 3 -d timescale of our integrations. By considering the strands as being composed of non-interacting test particles, we have also neglected any effects due to collisions between the particles and the self-gravity of the ring. However, by detecting and removing all particles that came within 50 km of the centre of our spherical satellite, we can monitor all collisions between the test particles and Prometheus.

Figure 2 shows six views of the evolved system. In all these plots, the consequences of previous passages of the satellite are clearly visible past Prometheus with distortions in the strands (particularly the inner one) and in the background distribution of particles. The structure in the ring is time-dependent even for those particles that have already encountered the satellite. This happens because, although the particles are no longer being perturbed, we are seeing the outcomes of the velocity kicks that were primarily induced in the various ring particles when the satellite was near its apoapse (mean anomaly, $M = 180^\circ$). Furthermore, the channels in the ring are almost undetectable in Fig. 2a and f but are striking in Fig. 2c and d when Prometheus is near apoapse. As shown in Fig. 4 of ref. 9, the largest kicks in a and e happen to particles that pass Prometheus near its apoapse, but the sign of Δa changes at this point; this means that channels will open in the rings every time Prometheus is near its apoapse.

Figure 2e and f best illustrates the formation of a streamer. The satellite is moving away from apoapse and, as seen in the bump of

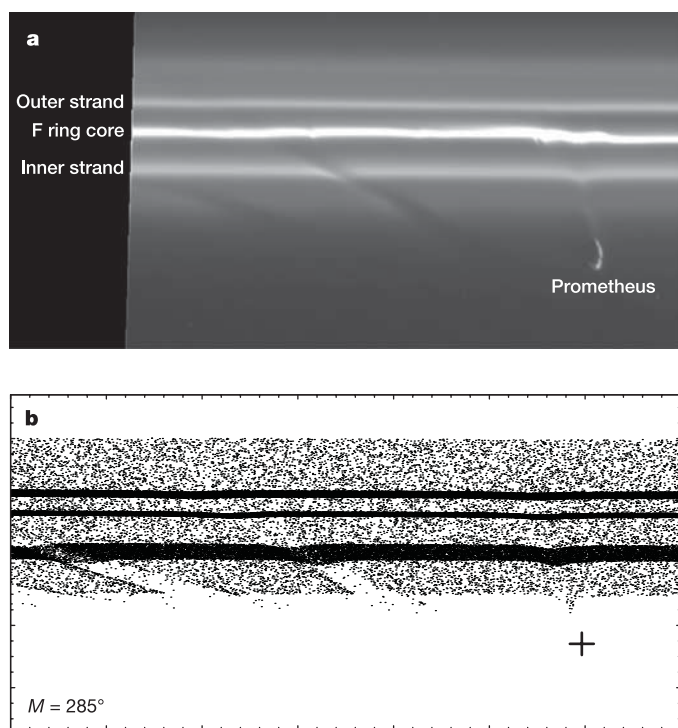


Figure 3 | Comparison between a Cassini image and our model for the same configuration. **a**, A reprojected Cassini image (N1477737741) of the F ring and Prometheus taken on 29 October 2004, and **b**, the corresponding numerical model when the satellite had a mean anomaly of 285° (intermediate between the configurations shown in Fig. 2e and f); the cross (+) denotes the position of Prometheus' centre. In each case the longitude range covers 7° and the radial range 1,500 km. The dark area at the left in **a** was outside the field of view of the ISS narrow angle camera. The existence of a streamer and sheared channels in both image and simulation is clearly evident. In this particular configuration, the model predicts a brightening of each channel on the side furthest from Prometheus, which agrees with the observed structure in the image.

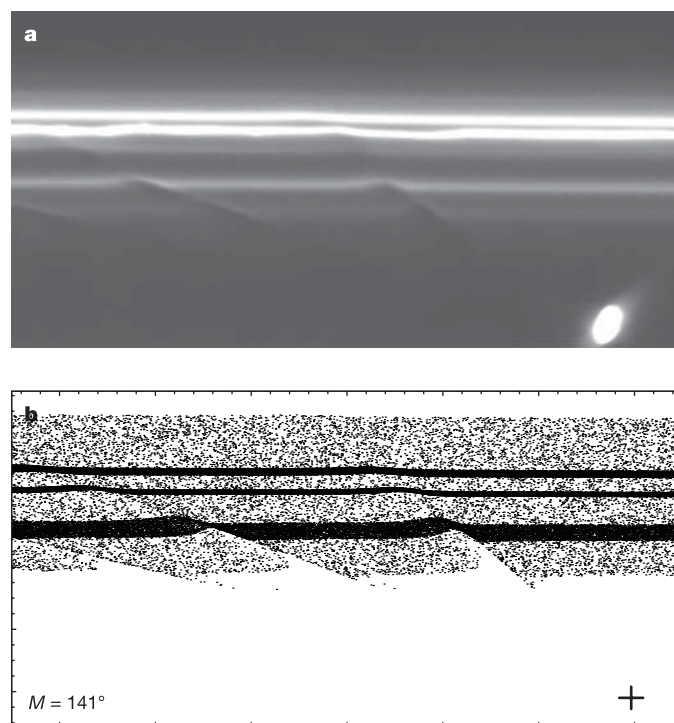


Figure 4 | Another comparison between a Cassini image and our model for the same configuration. **a**, As Fig. 3 but for N1492087204, taken on 13 April 2005, and **b**, as Fig. 3 but for a mean anomaly of 141° (intermediate between the configurations shown in Fig. 2b and c). Longitude range and radial range as Fig. 3. The Cassini image in **a** corresponds to the region at the right of the mosaic in Fig. 1d. There is excellent agreement between theory and observation. Note that the model predicts a brightening on the side of each channel that is closest to Prometheus (the reverse of the situation in Fig. 3), just as observed in the corresponding image.

particles immediately to the right of Prometheus, appears to be dragging particles out of the ring. Our integrations have shown that these particles originate in the innermost strand and background population. Note that the perturbations to create a channel are produced during the same pass but the channel is not fully developed until almost one orbital period later (Fig. 2d). In the meantime, because of the difference in semimajor axis across the ring, differential orbital motion has taken place and the channel has undergone keplerian shear. Note that a channel is not created by particles colliding with the satellite (in the configuration modelled here Prometheus does not enter the F ring), but is due instead to the forced change in orbital elements. However, the satellite's approach does perturb the adjacent particles but the effect is only seen later in the cycle. The integrations show that the particles in the streamer undergo a periodic departure from the ring caused by their forced eccentricity. Their keplerian motion returns them to the ring only for them to reappear periodically later in each cycle.

Figures 3 and 4 show direct comparisons between our model and two Cassini images showing Prometheus near the F ring, including one where a streamer is present. In each case there is excellent agreement, even at the level of the detailed structure on either side of each channel (compare Fig. 1b). It is also worthwhile considering what happens after the ring particles have been perturbed by Prometheus. The particles in a collisionless model continue in unperturbed orbits that are, of course, different from their earlier paths. The image mosaic shown in Fig. 1d confirms this concept. The continual shearing of the channels is evident, and the net result gives a rope-like appearance to the inner strands, reminiscent of features seen in the high-resolution SOI images of the edges of the Encke gap¹. The wavelength of such structures is $3\pi\delta a$, where δa is the difference in semimajor axis between the perturber (Prometheus) and the ring strand¹⁴. The different lengths between the edges of successive channels (shorter on the edge closer to the satellite versus longer where the channel meets the inner strand) simply reflects the different distance of δa . Figure 1d hints that the gradient along a channel is not constant; this was also detected in the SOI images of the F ring¹ taken by Cassini's Narrow Angle Camera (NAC). One of the predictions of a simplified analytical model⁹ of the system is that the size of the change in the orbital elements will be a function of the particle's eccentricity, and so we may be seeing the consequence of different eccentricities in different strands.

Our model is only a first step. It takes no account of collisions between particles or any self-gravity in the ring and yet these processes should be important for the F ring's optically thicker core. Nor does it account for the observed multi-strand nature of the ring^{1,8} or the existence of discrete clumps. Nevertheless, it provides a plausible mechanism for the origin of intricate structures detected in the F ring, and suggests that streamers, channels and a variety of other phenomena¹ can all be understood in terms of the simple gravitational effect of a satellite on a ring particle. Our model predicts that the detailed structure of the streamers and channels will

vary periodically (while still undergoing keplerian shear) on a time-scale of one orbital period, and this can be tested with future Cassini observations. We have also carried out additional simulations, which show that the exact geometry of the streamers is sensitive to the chosen mass of Prometheus and how close it gets to the F ring. Therefore we can use our model and Cassini images to provide an independent method for determining the mass of Prometheus^{5,13}.

Finally, we note that differential precession caused by Saturn's oblateness is driving Prometheus towards ever-deeper incursions into the F ring with more extreme perturbations, which will culminate in early December 2009 when the apoapse of Prometheus is aligned with the periapse of the F ring. In this configuration we anticipate that Prometheus will suffer impacts from F ring particles and that new structures will be detected in the F ring.

Received 16 July; accepted 2 September 2005.

1. Porco, C. C. *et al.* Cassini imaging science: Initial results on Saturn's rings and small satellites. *Science* **307**, 1226–1236 (2005).
2. Borderies, N. & Goldreich, P. The variations in eccentricity and apse precession rate of a narrow ring perturbed by a close satellite. *Icarus* **53**, 84–89 (1983).
3. Kolvoord, R. A., Burns, J. A. & Showalter, M. R. Periodic features in Saturn's F ring. *Nature* **345**, 695–697 (1990).
4. Murray, C. D. & Giuliatti Winter, S. M. Periodic collisions between the moon Prometheus and Saturn's F ring. *Nature* **380**, 139–141 (1996).
5. Jacobson, R. A. & French, R. G. Orbits and masses of Saturn's coorbital and F-ring shepherding satellites. *Icarus* **172**, 382–387 (2004).
6. Bosh, A., Olkin, C. B., French, R. G. & Nicholson, P. D. Saturn's F ring: Kinematics and particle sizes from stellar occultation studies. *Icarus* **157**, 57–75 (2002).
7. Porco, C. C. *et al.* Cassini imaging science: Instrument characteristics and anticipated scientific investigations at Saturn. *Space Sci. Rev.* **115**, 363–497 (2004).
8. Murray, C. D., Gordon, M. K. & Giuliatti Winter, S. M. Unravelling the strands of Saturn's F ring. *Icarus* **129**, 304–316 (1997).
9. Showalter, M. R. & Burns, J. A. A numerical study of Saturn's F ring. *Icarus* **52**, 526–544 (1982).
10. Kolvoord, R. A. & Burns, J. A. Three-dimensional perturbations of particles in a narrow planetary ring. *Icarus* **95**, 253–264 (1992).
11. Giuliatti Winter, S. M. *The Dynamics of Saturn's F Ring*. PhD thesis, Queen Mary and Westfield College, Univ. London (1994).
12. Giuliatti Winter, S. M., Murray, C. D. & Gordon, M. K. Perturbations to Saturn's F ring strands at their closest approach to Prometheus. *Planet. Space Sci.* **48**, 817–827 (2000).
13. Renner, S., Sicaudy, B. & French, R. G. Prometheus and Pandora: masses and orbital positions during the Cassini tour. *Icarus* **174**, 230–240 (2005).
14. Dermott, S. F. in *Planetary Rings* (eds Greenberg, R. & Brahic, A.) 589–637 (Univ. Arizona Press, Tucson, 1984).

Acknowledgements C.D.M., K.B., N.C. and M.W.E. are grateful to the UK Particle Physics and Astronomy Research Council for financial support. C.C. thanks the Astronomy Unit, Queen Mary, University of London and the Mexican Council for Science and Technology, CONACYT, for financial support. J.A.B. was supported by the Cassini project and Cornell University. C.C.P. acknowledges support from NASA/JPL and the Cassini project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.D.M. (C.D.Murray@qmul.ac.uk).

LETTERS

Observation of spin Coulomb drag in a two-dimensional electron gas

C. P. Weber¹, N. Gedik^{1,2}, J. E. Moore¹, J. Orenstein¹, J. Stephens³ & D. D. Awschalom³

An electron propagating through a solid carries spin angular momentum in addition to its mass and charge. Of late there has been considerable interest in developing electronic devices based on the transport of spin that offer potential advantages in dissipation, size and speed over charge-based devices¹. However, these advantages bring with them additional complexity. Because each electron carries a single, fixed value ($-e$) of charge, the electrical current carried by a gas of electrons is simply proportional to its total momentum. A fundamental consequence is that the charge current is not affected by interactions that conserve total momentum, notably collisions among the electrons themselves². In contrast, the electron's spin along a given spatial direction can take on two values, $\pm\hbar/2$ (conventionally $\uparrow, \downarrow$), so that the spin current and momentum need not be proportional. Although the transport of spin polarization is not protected by momentum conservation, it has been widely assumed that, like the charge current, spin current is unaffected by electron–electron (e – e) interactions. Here we demonstrate experimentally not only that this assumption is invalid, but also that over a broad range of temperature and electron density, the flow of spin polarization in a two-dimensional gas of electrons is controlled by the rate of e – e collisions.

In this work, spin diffusion is characterized by the transient spin grating technique³, which is based on the optical injection of spin-polarized electrons. The two-dimensional electron gas (2DEG) resides in a GaAs quantum well, in which the carriers are donated by Si impurities doped into the GaAlAs barrier layers. Near-bandgap illumination of the GaAs excites electrons whose initial spin is determined by the helicity of the light⁴. If the GaAs is excited by two non-collinear, coherent beams of light with orthogonal linear polarization, then in the region where the beams interfere the helicity varies sinusoidally from plus one to minus one. The optical-helicity wave generates a wave of electron-spin polarization with the same spatial frequency, which in turn generates a sinusoidal variation (grating) in the index of refraction through the Kerr effect. The wavevector of the injected spin-density wave is in the plane of the 2DEG, and the spin polarization is oriented perpendicular to this plane.

The time evolution of the transient spin grating directly reveals the nature of spin transport and relaxation in the electronic system, functioning like a time-domain version of neutron scattering. We measure the spin polarization by detecting the diffraction of a probe beam off the grating. A sensitive coherent detection scheme (Methods) enabled acquisition of the ~ 150 grating decays required to characterize the spin dynamics for each sample throughout the temperature–wavevector (T – q) parameter space.

In this Letter we present results for three quantum well samples, with electron concentrations of 7.8×10^{11} , 4.3×10^{11} and

$1.9 \times 10^{11} \text{ cm}^{-2}$, corresponding to Fermi temperatures of 400, 220 and 100 K, respectively. Figure 1 shows the initial decay rate of the spin grating as function of T in the most heavily doped sample, for several grating wavevectors from 0.4×10^4 to $2.5 \times 10^4 \text{ cm}^{-1}$. The dependence on T can be described in terms of three regions. For $100 \text{ K} < T < 300 \text{ K}$, the decay rate varies slowly. For $50 \text{ K} < T < 100 \text{ K}$, the decay rate increases rapidly with decreasing T , and for $T < 50 \text{ K}$, it reaches a slowly varying plateau.

We begin by discussing the decay rate where it varies slowly, that is, below 50 K and above 100 K. In the high- T region, the spin dynamics

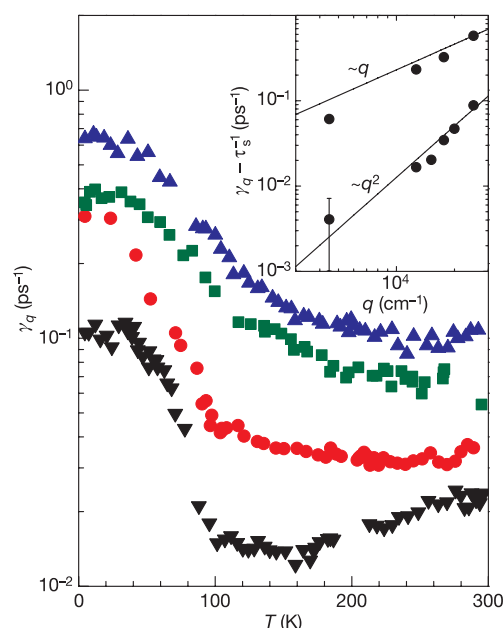


Figure 1 | Spin-grating decay at various wavevectors (q) and temperatures (T) for the sample with Fermi temperature 400 K. Main panel, the initial decay rate, γ_q , of the spin grating as a function of T for (bottom to top) $q = 0.45 \times 10^4$, 1.3×10^4 , 1.8×10^4 and $2.5 \times 10^4 \text{ cm}^{-1}$. Inset, the initial decay rate of the spin grating as a function of q . Points are $\gamma_q - \tau_s^{-1}$; τ_s is obtained from decay of homogenous ($q = 0$) spin excitation. Error bars (s.d.) are the size of the points except as shown. Lower points and line, room temperature. The line is a fit of the data to $\gamma_q = \tau_s^{-1} + D_s q^2$, giving a spin diffusion length $L_s = (D_s \tau_s)^{1/2} = 0.81 \mu\text{m}$ and a ‘spin mean-free-path’ $l = 2D_s/v_F = 60 \text{ nm}$. The observation of diffusive motion is internally consistent, as l is much smaller than both L_s and the smallest grating wavelength, $2.5 \mu\text{m}$. Upper points and line, 5 K. The line has slope = 1, corresponding to ballistic, rather than diffusive, spin-motion with a velocity of $2.3 \times 10^7 \text{ cm s}^{-1}$.

¹Physics Department, University of California, Berkeley, and Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ²Laboratory for Molecular Sciences, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, California 91125, USA. ³Center for Spintronics and Quantum Computation, University of California, Santa Barbara, California 93106, USA.

can be accurately described in terms of independent processes of spin diffusion and spin relaxation. In this description, the decay rate varies with q quadratically, as $\gamma_q = \tau_s^{-1} + D_s q^2$, where D_s is the spin diffusion coefficient and τ_s is the spin relaxation time³. In the inset to Fig. 1 we plot $\gamma_q - \tau_s^{-1}$ versus q at 295 K (lower points) and 5 K (upper points), on logarithmic axes. Here $1/\tau_s$ is independently determined from the decay rate of the circular dichroism induced by a circularly polarized pump beam⁵ (see Supplementary Information). A comparison of the 295 K data with a line of slope two shows that the decay of the grating is well described by diffusive dynamics, with $D_s = 130 \text{ cm}^2 \text{ s}^{-1}$ and $\tau_s = 50 \text{ ps}$.

Next, we examine the spin-grating dynamics at $T < 50 \text{ K}$. As shown in Fig. 1 inset, the initial decay rates at 5 K are linear in q at the higher wavevectors. The change in power law exponent from two to one indicates that a crossover from diffusive to ballistic dynamics takes place as the sample temperature is lowered. In the ballistic regime, electrons propagate a distance comparable to the grating wavelength, Λ , without scattering and the initial decay rate is $v_F q$, the reciprocal of the time required for an electron moving with the Fermi velocity (v_F) to traverse a distance $\Lambda/2\pi$.

Although the grating's initial decay rate saturates near $v_F q$ when T reaches $\sim 50 \text{ K}$, its time dependence continues to change as T is lowered further. Figure 2 shows the grating amplitude as function of time for several temperatures between 5 K and 100 K, measured with a grating wavevector of $2.5 \times 10^4 \text{ cm}^{-1}$ (the T indicated is the lattice temperature, which is below the electron temperature, as will be discussed later). An oscillatory structure appears in the decay curves, becoming increasingly pronounced as T decreases. The growth of these oscillations is a consequence of the increase of the mean-free-path, l , in the regime where $ql \geq 1$.

To determine D_s from data such as those in Fig. 2, we use an expression for the time dependence of a spin fluctuation that is applicable throughout the diffusive–ballistic crossover regime. If a spin polarization wave is introduced at $t = 0$, its subsequent time dependence is the Fourier transform of $S(q, \omega) \propto [i\omega - D(q, \omega)q^2]^{-1}$, where $D(q, \omega)$ is the dynamic spin diffusivity. In the limit $q \ll k_F$,

$$D(q, \omega) = \frac{v_F/2}{\sqrt{(i\omega/v_F - 1/l)^2 + q^2}} \quad (1)$$

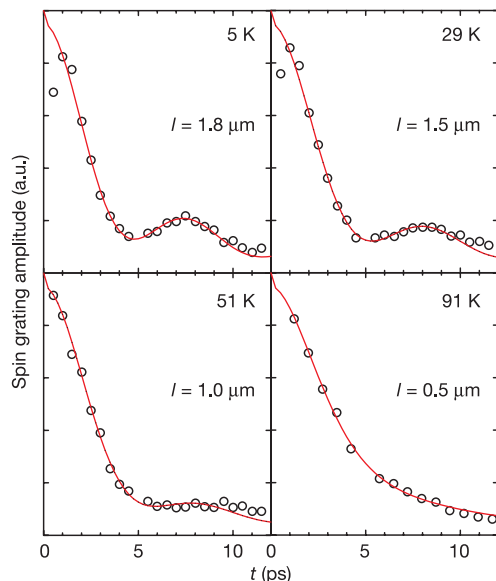


Figure 2 | Time dependence of the spin grating's amplitude. The lines are fits of the data to $S(q, \omega)$. The values of l determined from these fits are indicated in each panel. Owing to laser heating, the temperature T_e of the electron gas is higher than the lattice temperatures indicated.

where equation (1) extrapolates from the small- q limit⁶ to the ballistic regime. In attempting to fit the grating decay curves in the plateau regime, we found that equation (1) is not quite sufficient to describe the data. It is necessary to add to the Fourier transform of $S(q, \omega)$ a small, slowly decaying exponential with relative initial amplitude ~ 0.1 and characteristic time $\sim 25 \text{ ps}$. We speculate that this slow exponential may originate from a small fraction of localized electrons. The solid lines through the data points in Fig. 2 show the results of the fitting procedure, with fitting parameters l , v_F and the amplitude and time constant of the slow exponential. Despite the complicating presence of the slow exponential, we believe that the fits give an accurate indication of l , as this is the only parameter that determines the rate at which the oscillations are damped. Finally, the spin diffusion coefficient is determined from the relation $D_s = v_F l/2$.

The temperature dependence of D_s obtained from our analysis of the spin-grating dynamics is shown in Fig. 3 for quantum wells of different electron density. For the two lower-density samples (middle and lower panels), the dynamics were diffusive at all T , consistent with their lower mobility. To characterize charge transport in the same set of samples, we performed 4-probe measurements of the 2D charge conductance, σ_c , carrier density, n , and mobility, μ , on chips from the same set of wafers. Together, these measurements allow us to test the assumption that the scattering processes that control spin diffusion and charge conduction are the same. The link between conductance and diffusion coefficient is the Einstein relation, $D_s = \sigma_s/e^2 \chi_s$, where σ_s and χ_s are the spin conductance and

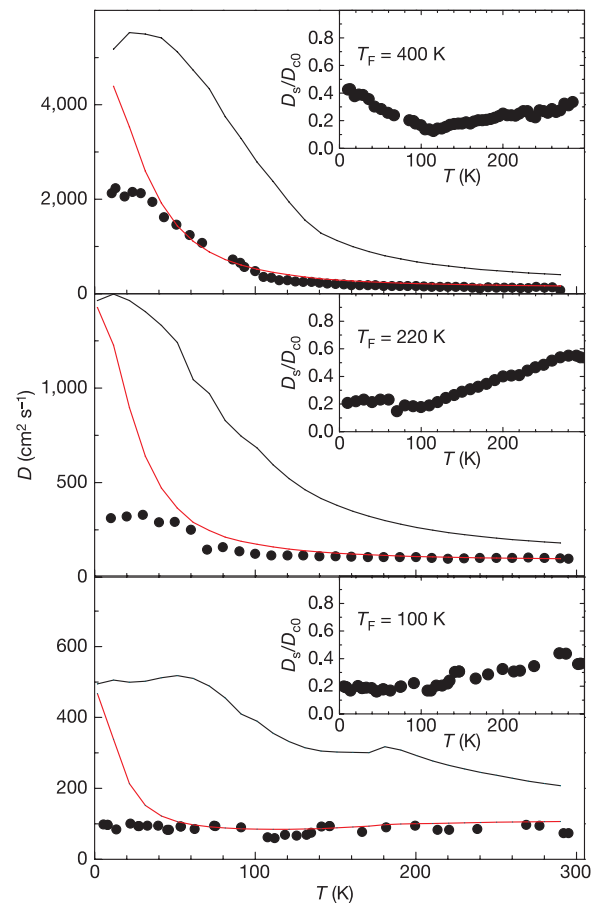


Figure 3 | Comparison of motion of spin and charge, for samples with the Fermi temperatures (T_F) shown. Dots (main panels), spin-diffusion coefficients D_s determined from optical measurements. Black lines, quasiparticle diffusion coefficients D_{c0} determined from transport data. Insets, D_s/D_{c0} . Red lines, D_s predicted from spin Coulomb drag theory, taking $\chi_s = \chi_0$.

susceptibility, respectively. If the spin and charge scattering rates were the same (that is, $\sigma_c = \sigma_s$), then D_s would equal $(\chi_0/\chi_s)D_{c0}$ (ref. 7), where $D_{c0} = \sigma_c/e^2\chi_0$ and $\chi_0 = N_F(1 - e^{-E_F/k_B T})$ is the non-interacting susceptibility (see Supplementary Information; E_F is the Fermi energy, N_F is the density of states at E_F and k_B is Boltzmann's constant). Physically, D_{c0} is the quasiparticle diffusion coefficient⁷, approaching $\mu E_F/e$ and $\mu k_B T/e$ in the degenerate and non-degenerate regimes, respectively. D_{c0} calculated from the 4-probe transport data and plotted in Fig. 3 is considerably larger than D_s at all T and for each of the samples. The ratio is far greater than can be accounted for by many-body enhancement of the spin susceptibility, as the factor χ_s/χ_0 is less than 1.4 in this range of electron density^{8,9}.

The contrast in the diffusion coefficients of charge and spin is surprising, as the assumption $D_s = D_{c0}$ is widely used in modelling spin transport in semiconductors. However, this assumption fails to take into account e - e collisions, whose rate can be much faster than those of impurity or phonon scattering. The e - e scattering events can be ignored in the description of charge transport because they conserve total momentum. However, they can have a profound effect on spin transport, as illustrated in Fig. 4. For the collision depicted between electrons with opposite spin, the charge current is conserved while the spin current reverses direction.

D'Amico and Vignale have proposed that the microscopic process shown in Fig 4 can change the nature of macroscopic spin transport. Seen macroscopically, e - e collisions transfer momentum between the spin-up and spin-down populations, creating a force damping their relative motion that these authors term 'spin Coulomb drag' (SCD)¹⁰. Spin diffusion, which requires a counterflow of the spin populations, is damped by SCD, while charge diffusion is not. (The recently observed^{11,12} spin Hall effect also involves the counterflow of spin populations, and so should be damped by SCD.) According to D'Amico and Vignale¹⁰, the reduction of D_s relative to D_{c0} is:

$$\frac{D_s}{D_{c0}} = \left(\frac{\chi_0}{\chi_s}\right) \frac{1}{1 + |\rho_{\uparrow\downarrow}|/\rho} \quad (2)$$

where $\rho = 1/\sigma_c$ is the charge resistance and $\rho_{\uparrow\downarrow}$ is the spin drag resistance, parameterizing the rate of momentum exchange between spin $\uparrow$ and $\downarrow$ electrons. D'Amico and Vignale, and Flensberg and Jensen¹³, have calculated $\rho_{\uparrow\downarrow}(T)$ for a 2DEG using the random phase approximation, obtaining results that depend only on the electron density of the quantum well.

Equation (2) predicts that despite the complex T dependences of the individual diffusion coefficients, their ratio depends primarily on the single factor, $|\rho_{\uparrow\downarrow}|/\rho$. We test this prediction in Fig. 5, without invoking any assumptions or adjustable parameters, by plotting D_{c0}/D_s (the inverse of equation (2)) versus $|\rho_{\uparrow\downarrow}|/\rho$ for each of the three samples measured in this study. The transport coefficients are taken directly from our measurements, while $\rho_{\uparrow\downarrow}$ was calculated using equation (2) of ref. 14. The resulting graph reveals the simple linear dependence of D_{c0}/D_s on $|\rho_{\uparrow\downarrow}|/\rho$ predicted by equation (2) over a

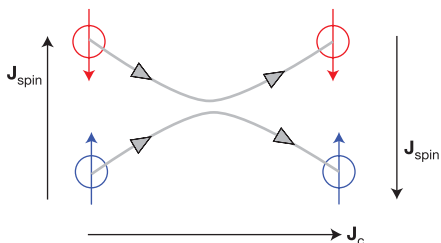


Figure 4 | A representation of e - e scattering that does not conserve spin-current. Before the collision the spin-current, J_{spin} , is positive; after, it is negative. The charge current, J_c , does not change. Colours correspond to spin states

large range of $|\rho_{\uparrow\downarrow}|/\rho$, implying that SCD is indeed the origin of the large suppression of D_s relative to D_{c0} . The fact that the slope is slightly greater than unity is consistent with the expectation that the many-body enhancement of χ_s relative to χ_0 is small in this density regime^{8,9}. Finally, the fact that D_{c0}/D_s extrapolates to near unity as $|\rho_{\uparrow\downarrow}|/\rho \rightarrow 0$ indicates that the spin and charge diffusion coefficients approach each other in the limit that the spin drag resistance becomes smaller than the ordinary resistance. This result provides independent evidence that the spin grating and four-probe techniques used in this work accurately measure equilibrium spin and charge transport coefficients, respectively.

Returning to the T dependence shown in Fig. 3, the solid lines show the prediction of equation (2) for D_s with the factor χ_0/χ_s set equal to unity. As could be anticipated from the discussion of Fig. 5, SCD quantitatively accounts for the suppression of D_s relative to D_{c0} over a broad range of temperature and electron density. It is clear, however, that the measured D_s consistently departs from theory below 40 K. We believe that this discrepancy indicates that at low T the photoexcited electron gas does not cool to the lattice temperature. If the electron gas retains the heat, Q , deposited by the excitation, its temperature T_e will rise to approximately $(T^2 + 2Q/\beta)^{1/2}$, where $\beta = 5.3 \times 10^5 \text{ eV cm}^{-2} \text{ K}^{-2}$ is the temperature coefficient of the electronic specific heat. We estimate $Q = 4.8 \times 10^8 \text{ eV cm}^{-2}$, assuming that each absorbed photon deposits approximately 10 meV (the energy width of the laser pulse) into the Fermi sea. The resulting estimate for the minimum T_e is indeed $\sim 35 \text{ K}$.

Finally, we note that SCD can be highly advantageous for spintronic applications, as it increases the distance that a spin packet can be dragged by an electric field, E , before it spreads owing to diffusion¹⁵. The length L_D that a packet of width w will drift before it broadens by a factor of two is $w^2 e \mu / D_s$. In the absence of SCD, the ratio μ/D_s equals $e/k_B T_F$ (where T_F is Fermi temperature) or $e/k_B T$ in the degenerate or non-degenerate regimes, respectively, and L_D/w is independent of the underlying scattering rates. In the degenerate regime, for example, $L_D/w = eEw/k_B T_F$; drifting a spin packet farther than w is only possible in a strong E limit, where the potential drop across the packet exceeds the Fermi energy. Introducing SCD slows the counterflow of spin $\uparrow$ and $\downarrow$ electrons without affecting their co-propagation, amplifying L_D/w by the factor $1 + |\rho_{\uparrow\downarrow}(T)|/\rho$. Clean materials with strong e - e scattering will have the largest values

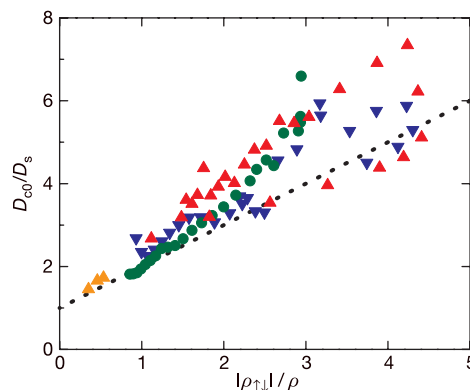


Figure 5 | Relation between suppression of spin diffusion and spin drag resistance. x -axis, ratio of $|\rho_{\uparrow\downarrow}(T)|$ (where $\rho_{\uparrow\downarrow}$ is the spin drag resistance), determined from SCD theory¹⁴, to measured resistivity, ρ . y -axis, ratio of quasiparticle diffusion coefficient, D_{c0} , to spin diffusion coefficient, D_s . Temperature is an implicit parameter. Points are for samples with $T_F = 400 \text{ K}$ (red), 220 K (green) and 100 K (blue). Points for $T < 40 \text{ K}$ are not shown because the electrons do not cool below 40 K. Orange points for a sample with $T_F = 400 \text{ K}$, with ρ increased by depositing a portion of the Si dopants into the well. Line has unity slope and intercept, indicating the prediction of equation (2) for $\chi_s = \chi_0$. For points above the line, $\chi_s > \chi_0$.

of $|\rho_{\parallel}(T)|/\rho$, and hence be the best media for propagation of spin information.

METHODS

Quantum well characteristics. The GaAs/Ga_{0.7}Al_{0.3}As samples were grown in the (100) direction by molecular beam epitaxy, and each consist of ten quantum wells of thickness 12 nm, separated by 48 nm barriers. The Si impurities were deposited in eight single atomic layers in the centre 14 nm of each barrier to maximize their distance from the 2DEG. The carrier concentration, n , mobility, μ , and electrical resistivity, ρ , were measured using 4-probe transport techniques without illumination. For the samples with n of 7.8×10^{11} , 4.3×10^{11} and $1.9 \times 10^{11} \text{ cm}^{-2}$ per quantum well, at low temperature μ reached 240,000, 92,000 and $69,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, respectively.

Optical methods. The two interfering beams that generate the optical-helicity wave derive from a Ti:sapphire laser, which produces a train of optical pulses with duration 100 fs, interpulse separation 11 ns and centre wavelength 820 nm. The incident power density for most measurements was $\sim 500 \text{ W cm}^{-2}$, corresponding to $\sim 6 \text{ W cm}^{-2}$ absorbed per quantum well. For $T > 35 \text{ K}$, grating decay rates did not change when measured at incident powers down to 100 W cm^{-2} , suggesting that photoinduced holes do not play a significant role in the electron spin transport (typical electron hole recombination times were $\sim 750 \text{ ps}$). At low T the grating decay rate increased slowly with decreasing power, consistent with the electron heating model described in the main text.

The grating wavevector was directed along the GaAs (001) direction. To detect the induced spin grating, we mixed diffracted and transmitted probes to produce a photodetector current linear in the diffracted field^{16–18}. The decisive advantage in this scheme is realized by modulating the relative phase of these two beams sinusoidally at 1.2 kHz. Synchronous detection with a lock-in amplifier at the modulation frequency leads to considerable rejection of laser noise and stray light.

Received 29 April; accepted 2 September 2005.

1. Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, Berlin, 2002).
2. Ziman, J. M. *Electrons and Phonons: The Theory of Transport Phenomena in Solids* (Oxford Univ. Press, New York, 2001).
3. Cameron, A. R., Riblet, P. & Miller, A. Spin gratings and the measurement of electron drift mobility in multiple quantum well semiconductors. *Phys. Rev. Lett.* **76**, 4793–4796 (1996).
4. Meier, F. & Zakharchenya, B. *Optical Orientation* (North-Holland, Amsterdam, 1984).
5. Bar-Ad, S. & Bar-Joseph, I. Exciton spin dynamics in GaAs heterostructures. *Phys. Rev. Lett.* **68**, 349–352 (1992).
6. Rammer, J. *Quantum Transport Theory* (Perseus Books, Reading, Massachusetts, 1998).
7. Castellani, C., DiCastro, C., Kotliar, G., Lee, P. A. & Strinati, G. Thermal conductivity in disordered interacting-electron systems. *Phys. Rev. Lett.* **59**, 477–480 (1987).
8. Yarlagadda, S. & Giuliani, G. F. Spin susceptibility in a two-dimensional electron gas. *Phys. Rev. B* **40**, 5432–5440 (1989).
9. Kwon, Y., Ceperley, D. M. & Martin, R. M. Quantum Monte Carlo calculation of the Fermi liquid parameters in the two-dimensional electron gas. *Phys. Rev. B* **50**, 1684–1694 (1994).
10. D'Amico, I. & Vignale, G. Spin diffusion in doped semiconductors: The role of Coulomb interactions. *Europhys. Lett.* **55**, 566–572 (2001).
11. Kato, Y. K., Myers, R. C., Gossard, A. C. & Awschalom, D. D. Observation of the spin Hall effect in semiconductors. *Science* **306**, 1910–1913 (2004).
12. Wunderlich, J., Kaestner, B., Sinova, J. & Jungwirth, T. Experimental observation of the spin Hall effect in a two-dimensional spin-orbit coupled semiconductor system. *Phys. Rev. Lett.* **94**, 047204 (2005).
13. Flensberg, K., Jensen, T. S. & Mortensen, N. A. Diffusion equation and spin drag in spin-polarized transport. *Phys. Rev. B* **64**, 245308 (2001).
14. D'Amico, I. & Vignale, G. Spin Coulomb drag in the two-dimensional electron liquid. *Phys. Rev. B* **68**, 045307 (2003).
15. Kikkawa, J. M. & Awschalom, D. D. Lateral drag of spin coherence in gallium arsenide. *Nature* **397**, 139–141 (1999).
16. Vohringer, P. & Scherer, N. F. Transient grating optical heterodyne detected impulsive stimulated Raman-scattering in simple liquids. *J. Phys. Chem.* **99**, 2684–2695 (1995).
17. Chang, Y. J., Cong, P. & Simon, J. D. Optical heterodyne-detection of impulsive stimulated Raman-scattering in liquids. *J. Phys. Chem.* **99**, 7857–7859 (1995).
18. Gedik, N. & Orenstein, J. Absolute phase measurement in heterodyne detection of transient gratings. *Opt. Lett.* **29**, 2109–2111 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. D'Amico and G. Vignale for sending us numerical evaluations of their integral expression for the spin drag resistance. This work was funded by the US DOE, DARPA, and NSFDMR. We also acknowledge support from the Fannie and John Hertz Foundation (C.P.W.) and the Hellman Foundation (J.E.M.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.P.W. (cpweber@lbl.gov).

LETTERS

Strong quantum-confined Stark effect in germanium quantum-well structures on silicon

Yu-Hsuan Kuo¹, Yong Kyu Lee¹, Yangsi Ge¹, Shen Ren¹, Jonathan E. Roth¹, Theodore I. Kamins^{1,2}, David A. B. Miller¹ & James S. Harris¹

Silicon is the dominant semiconductor for electronics, but there is now a growing need to integrate such components with optoelectronics for telecommunications and computer interconnections¹. Silicon-based optical modulators have recently been successfully demonstrated^{2,3}; but because the light modulation mechanisms in silicon⁴ are relatively weak, long (for example, several millimetres) devices² or sophisticated high-quality-factor resonators³ have been necessary. Thin quantum-well structures made from III-V semiconductors such as GaAs, InP and their alloys exhibit the much stronger quantum-confined Stark effect (QCSE) mechanism⁵, which allows modulator structures with only micrometres of optical path length^{6,7}. Such III-V materials are unfortunately difficult to integrate with silicon electronic devices. Germanium is routinely integrated with silicon in electronics⁸, but previous silicon–germanium structures have also not shown strong modulation effects^{9–13}. Here we report the discovery of the QCSE, at room temperature, in thin germanium quantum-well structures grown on silicon. The QCSE here has strengths comparable to that in III-V materials. Its clarity and strength are particularly surprising because germanium is an indirect gap semiconductor; such semiconductors often display much weaker optical effects than direct gap materials (such as the III-V materials typically used for optoelectronics). This discovery is very promising for small, high-speed¹⁴, low-power^{15–17} optical output devices fully compatible with silicon electronics manufacture.

Quantum wells are thin (for example, 10 nm) layers of semiconductors surrounded by barrier materials. Usually the barriers are chosen to confine electrons in the conduction band and holes (or absences of electrons) in the valence band inside the quantum well; this is called type-I band alignment. The QCSE gives strong spectral shifts of the optical absorption edge with applied electric field near the direct bandgap (that is, a band structure in which the energy minima for electrons and holes lie at the same momentum) in such type-I quantum wells¹⁸. Usually these relevant band minima are at zero momentum (the Γ point, or ‘zone centre’).

The QCSE is routinely used in high-performance quantum-well modulators for telecommunications. It possesses a number of attractive properties that allow for modulators with only micrometres of path length that can be incorporated into large arrays (for example, >38,000 devices) attached to complementary metal-oxide-semiconductor (CMOS) silicon circuits⁶, as well as devices at telecommunications wavelengths of $\sim 1.5\ \mu\text{m}$ (ref. 7). Waveguide QCSE devices typically have lengths of only $\sim 100\text{--}400\ \mu\text{m}$ (see, for example, ref. 14). Recent work¹⁹ has demonstrated devices in InGaAs/InP quantum wells without waveguides that exhibit useful modulation at telecommunications wavelengths for $<1\ \text{V}$ drive, and with the relaxed alignment required for practical packaging. QCSE

devices do not require carrier injection, and are typically operated as reverse-biased diodes. The resulting low power dissipation—of the order of 10 mW per channel—allows large arrays of optical interconnects operating at high data rates from silicon chips^{15–17}. Theoretically, the QCSE is thought to operate at subpicosecond times^{20,21}, and devices with $>50\ \text{GHz}$ modulation bandwidth have been demonstrated¹⁴.

Unlike the III-V compounds typically used for QCSE modulators, both silicon and germanium have indirect lowest-energy bandgaps (meaning that the electron and hole energy minima have different momenta). Previously, type-I SiGe/Si quantum wells have shown no or inefficient QCSE^{9–11}, whereas SiGe/Si quantum wells¹² and Ge/Si quantum dots¹³ (both of which have type-II band alignment in which the electrons and holes have minimum energy in different material layers) can exhibit large shifts of optical transitions with electric field, but have low absorption efficiency. Here we demonstrate clear quantum confinement effects in the optical absorption spectra of Ge quantum wells, associated with zone centre transitions, and in addition a clear and strong QCSE electroabsorption. Though germanium’s lowest bandgap is indirect, we exploit its direct bandgap at $\sim 0.8\ \text{eV}$ at room temperature; the band structure associated with this direct bandgap is qualitatively identical to that in III-V QCSE materials. There is also indirect absorption present at the same and

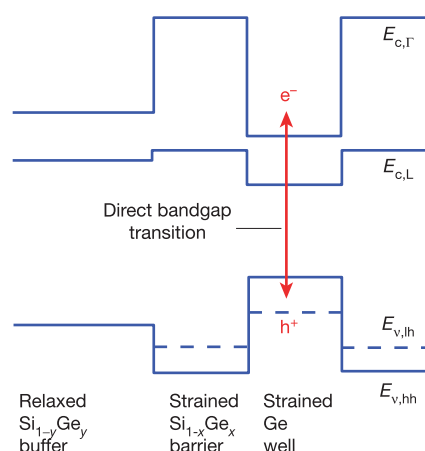


Figure 1 | The bandgap structure of a Ge/SiGe quantum well (not to scale). $E_{c,\Gamma}$ and $E_{c,L}$ are the energies of the bottom of the conduction band at zone centre (Γ point) and at the L valleys, respectively. $E_{v,lh}$ and $E_{v,hh}$ are the energies of the tops of the light-hole and heavy-hole valence bands, respectively. The Ge/Si_{0.15}Ge_{0.85} quantum well on relaxed Si_{0.1}Ge_{0.9} has type-I alignment at the zone centre and quantum-confines carriers inside the Ge well.

¹Solid State and Photonics Laboratory, Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA. ²Quantum Science Research, Hewlett-Packard Laboratories, Palo Alto, California 94304, USA.

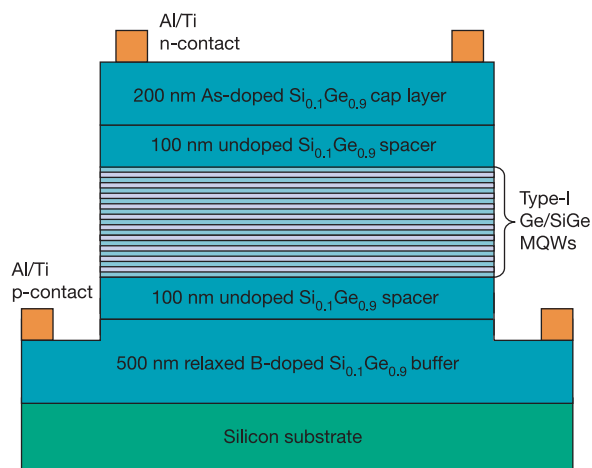


Figure 2 | Schematic diagram of a p-i-n diode. The cross-sectional view shows the structure of strained Ge/SiGe multiple quantum wells (MQWs) grown on silicon on relaxed SiGe direct buffers (not to scale). In the measurements, light from a monochromator is incident on the top surface (that is, in a 'surface-normal' configuration) on the open area inside the rectangular frame top electrode (that is, between the portions of the AlTi n-contacts shown in this cross-section).

lower photon energies, but it is much weaker, allowing the direct optical absorption to dominate.

We use strain-balanced Ge/SiGe multiple quantum wells (MQWs) grown on a relaxed Ge-rich SiGe buffer on silicon, giving type-I alignment at the Γ point. In strain-balanced structures, the average silicon concentration in the Ge/SiGe MQW layers equals that of the buffer layer, allowing the growth of thick structures. Figure 1 shows the resulting bandgap alignment. The band discontinuities of the heavy hole, light hole, and electron at the Γ point between our wells and barriers are calculated to be 101 meV, 47 meV, and 400 meV respectively, based on refs 22–24 with linear interpolation of the direct bandgap of SiGe between Si (to the Γ -2' band) and Ge. Note that this 400 meV Γ -point band discontinuity provides strong electron quantum confinement in the conduction band. There are, of course, lower-energy conduction band minima (L valleys) in the SiGe barriers. For quantum confinement, however, we expect the relevant bands in the barriers to be those with similar unit cell symmetries, as we expect that it is only to such bands that strong tunnelling can exist. Hence, we consider here the bands at the Γ point in the barriers for the calculation of quantum confinement effects. Because the Γ point in Ge wells is higher than the L valleys in the SiGe barriers, we do however expect that, even though the Γ point electron quantum well is relatively deep, electrons generated in the Ge wells will be rapidly scattered out into those L valleys where they can be swept out by electric fields (though not so fast as to prevent the quantum confinement exploited in optical absorption).

Figure 2 shows the Ge/SiGe MQW p-i-n diode structure. The layers are grown sequentially in a commercially available, single-wafer, cold-wall, reduced-pressure, chemical vapour deposition (RPCVD) reactor, using hydrogen carrier gas and silane and germane precursor gases. We use four-inch, boron-doped, (001)-oriented silicon wafer substrates with resistivity 10–20 Ω cm. After an *in situ* high-temperature cleaning, the layers are grown at 400 °C, with annealing at higher temperatures. Two 250-nm $\text{Si}_{0.1}\text{Ge}_{0.9}$ films doped with $5 \times 10^{18} \text{ cm}^{-3}$ boron atoms are grown sequentially and annealed at 850 °C for 30 min and 700 °C for 5 min, respectively, to reduce dislocations caused by the lattice mismatch and to form relaxed p-type buffer layers. A 100 nm undoped $\text{Si}_{0.1}\text{Ge}_{0.9}$ spacer layer is grown, followed by ten pairs of MQWs (10 nm Ge well/16 nm $\text{Si}_{0.15}\text{Ge}_{0.85}$ barrier) and another 100 nm undoped $\text{Si}_{0.1}\text{Ge}_{0.9}$ spacer layer. Finally, the structure is capped by an arsenic-doped $\text{Si}_{0.1}\text{Ge}_{0.9}$

layer with a doping level of 10^{19} cm^{-3} . Each wafer is then patterned by standard lithography and dry-etched to form square mesa structures with widths ranging from 200 to 1,400 μm . Al/Ti metal is evaporated, lifted off, and rapid-thermal-annealed to form rectangular frame ohmic contacts. Because the top As-doped (n-type) SiGe layer is expected to be strongly conducting, the electric field should be essentially perpendicular to the quantum wells throughout the structure. All materials, processing equipment and temperatures used are compatible with the processes used in fabricating CMOS silicon electronics.

We measure photocurrent spectra at room temperature for different diode reverse bias voltages using a chopped quartz-tungsten-halogen light bulb source with a 0.25 m monochromator with a 600 lines mm^{-1} grating and a 400 μm slit, and a lock-in amplifier. The light is incident normal to the surface, with random polarization in the surface plane. As in other quantum-well structures, we expect light- and heavy-hole transitions both to be allowed in this configuration. The responsivity (current per unit optical power) inside the device is deduced from the light intensity incident upon the open area of the mesa surface, correcting for surface reflection. The corresponding effective absorption coefficient spectra (Fig. 3) are calculated assuming one electron of current for every absorbed photon, as is common in fully depleted p-i-n diodes. (Presuming that there is no avalanche gain, this assumption at worst underestimates the absorption. The independence of the photocurrent on bias at high photon energies (for example, 0.94 eV) suggests no bias dependence of the number of electrons per photon, and hence no avalanching.) Clear quantum confinement is seen, with strong exciton peaks that we assign to electron-to-heavy-hole (e-hh; ~ 0.88 eV at 0 V) and electron-to-light-hole (e-lh; ~ 0.91 eV at 0 V) transitions. The effective absorption coefficient at the e-hh exciton peak, calculated on the basis of the total thickness of the wells and the barriers ($\sim 0.26 \mu\text{m}$), is $6,320 \text{ cm}^{-1}$. The half-width at half-maximum of the heavy-hole exciton peak is only ~ 8 meV, and the heavy-hole exciton peak is still clearly resolved even at $8 \times 10^4 \text{ V cm}^{-1}$ (3 V bias), indicating that the electric field is uniform inside the structure, which in turn implies very low background doping in the intrinsic region. The absorption edge at zero bias is ~ 80 meV higher than that of bulk unstrained Ge; this shift is near the calculated sum from quantum-well confinement (56 meV) and strain-induced shifts (36 meV)²³. As the quantum confinement energy originates primarily from the electron, the clarity of this quantum shift shows that the (quantum-mechanical) confinement at the Γ point is strong,

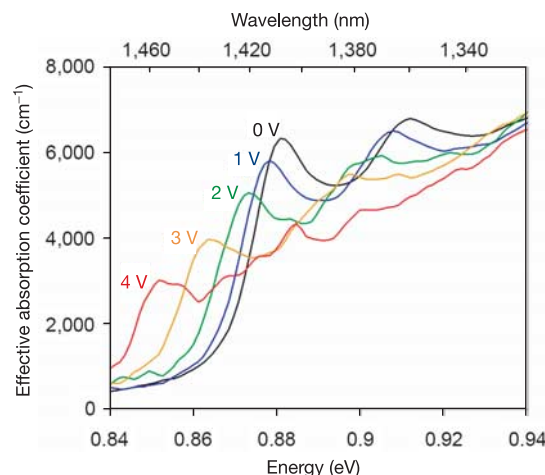


Figure 3 | Effective absorption coefficient spectra. Strong QCSE is observed at room temperature with reverse bias from zero to 4 V. The thickness for effective absorption coefficient calculations is based on the combination of Ge well and SiGe barrier thicknesses.

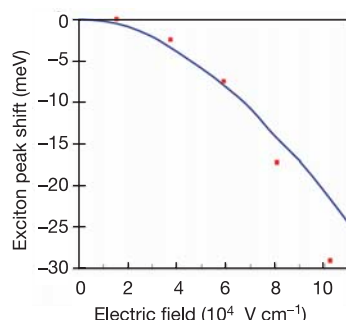


Figure 4 | Shifts of exciton peaks. Comparison of heavy-hole exciton peak shift from measurements (filled squares) and tunnelling resonance calculations (sum of electron and hole level shifts) (line).

despite the possibility of scattering to the lower indirect valleys.

With a reverse bias from 0 to 4 V, both peaks are red shifted by the QCSE. The e-hh exciton is shifted from 1,408 nm (0 V) to 1,456 nm (4 V). The maximum effective absorption coefficient change is $2,800 \text{ cm}^{-1}$ at 1,438 nm under 3 V bias. This is, to our knowledge, the first efficient electro-absorption modulation observed in group-IV materials, and its performance is comparable to high-quality (direct gap) III-V materials at similar wavelengths (see, for example, ref. 19). The clarity of the exciton peaks in the presence of a field is actually better than that of typical III-V structures at such wavelengths¹⁹, and the electroabsorption shows much clearer shifts than previous electroabsorption measurements in indirect III-V materials²⁵. With a 4 V bias, the absorption coefficient contrast is greater than 3 over a bandwidth ranging from 1,443 to 1,471 nm, with a peak value of 4.69 at 1,461 nm.

The possibility of operating different quantum-well designs at, say, 1,550 nm, which is compatible with long-distance telecommunications, will be the subject of future work. We also anticipate that waveguide modulator structures will be realizable using appropriate materials for waveguide cladding layers.

The measured shift agrees with simulated results (Fig. 4) obtained via the tunnelling resonance method^{5,18}. In addition, we evaluated the exciton binding energy shift as in ref. 18, using numerically evaluated electron and hole wavefunctions, though this correction is $<1 \text{ meV}$ and is neglected here. We used a Γ -valley electron effective mass of $0.041m_0 + 0.115(1-x)m_0$, and a heavy-hole effective mass of $0.28m_0 + 0.21(1-x)m_0$, where m_0 is the free electron mass and x is the Ge concentration^{26,27} (the relevant silicon (001) hole mass is based on Luttinger parameters²⁸).

We have demonstrated efficient QCSE in silicon-based structures, using strained Ge MQWs. The behaviour of the exciton peaks, the band edge shift and the shift in absorption coefficient are comparable to those observed in III-V materials at similar wavelengths. Our materials and fabrication processes are completely CMOS compatible and suitable for mass production. This approach is therefore very promising for silicon-based electro-absorption modulators operating at high speed, low power, low operating voltage and with small device areas.

Received 8 July; accepted 5 September 2005.

1. Miller, D. A. B. Rationale and challenges for optical interconnects to electronic chips. *Proc. IEEE* **88**, 728–749 (2000).
2. Liu, A. *et al.* A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor. *Nature* **427**, 615–618 (2004).

3. Xu, Q., Schmidt, B., Pradhan, S. & Lipson, M. Micrometre-scale silicon electro-optic modulator. *Nature* **435**, 325–327 (2005).
4. Soref, R. A. & Bennett, B. R. Electrooptical effects in silicon. *IEEE J. Quant. Electron.* **23**, 123–129 (1987).
5. Miller, D. A. B. *et al.* Band-edge electroabsorption in quantum well structures: the quantum-confined Stark effect. *Phys. Rev. Lett.* **53**, 2173–2176 (1984).
6. Arad, U. *et al.* Development of a large high-performance 2-D array of GaAs-AlGaAs multiple quantum-well modulators. *IEEE Photon. Tech. Lett.* **15**, 1531–1533 (2003).
7. Liu, C. P. *et al.* Design, fabrication and characterisation of normal-incidence 1.56- μm multiple-quantum-well asymmetric Fabry-Perot modulators for passive picocells. *IEICE Trans. Electron.* **E 86C**, 1281–1289 (2003).
8. Cressler, J. D. SiGe HBT technology: a new contender for Si-Based RF and microwave circuit applications. *IEEE Trans. Microwave Theory Tech.* **46**, 572–589 (1998).
9. Qasaimeh, O., Bhattacharya, P. & Croke, E. T. SiGe-Si quantum-well electroabsorption modulators. *IEEE Photon. Tech. Lett.* **10**, 807–809 (1998).
10. Miyake, Y., Kim, J. Y., Shiraki, Y. & Fukatsu, S. Absence of Stark shift in strained $\text{Si}_{1-x}\text{Ge}_x/\text{Si}$ type-I quantum wells. *Appl. Phys. Lett.* **68**, 2097–2099 (1996).
11. Li, C. *et al.* Observation of quantum-confined Stark shifts in SiGe/Si type-I multiple quantum wells. *J. Appl. Phys.* **87**, 8195–8197 (2000).
12. Park, J. S., Karunasiri, R. P. G. & Wang, K. L. Observation of large Stark shift in $\text{Ge}_{0.5}\text{Si}_{0.5}/\text{Si}$ multiple quantum wells. *J. Vac. Sci. Technol. B* **8**, 217–220 (1990).
13. Yakimov, A. I. *et al.* Stark effect in type-II Ge/Si quantum dots. *Phys. Rev. B* **67**, 125318 (2003).
14. Lewen, R., Irmscher, S., Westergren, U., Thylen, L. & Eriksson, U. Segmented transmission-line electroabsorption modulators. *J. Lightwave Technol.* **22**, 172–179 (2004).
15. Krishnamoorthy, A. V. & Miller, D. A. B. Scaling optoelectronic-VLSI circuits into the 21st century: a technology roadmap. *IEEE J. Select. Top. Quant. Electron.* **2**, 55–76 (1996).
16. Kibar, O., Van Blerkom, D. A., Fan, C. & Esener, S. C. Power minimization and technology comparisons for digital free-space optoelectronic interconnections. *J. Lightwave Technol.* **17**, 546–555 (1999).
17. Cho, H., Kapur, P. & Saraswat, K. C. Power comparison between high-speed electrical and optical interconnects for interchip communication. *J. Lightwave Technol.* **22**, 2021–2033 (2004).
18. Miller, D. A. B. *et al.* Electric field dependence of optical absorption near the bandgap of quantum well structures. *Phys. Rev. B* **32**, 1043–1060 (1985).
19. Helman, N. C., Roth, J. E., Bour, D. P., Altug, H. & Miller, D. A. B. Misalignment-tolerant surface-normal low-voltage modulator for optical interconnects. *IEEE J. Select. Top. Quant. Electron.* **11**, 338–342 (2005).
20. Schmitt-Rink, S., Chemla, D. S., Knox, W. H. & Miller, D. A. B. How fast is excitonic electroabsorption? *Opt. Lett.* **15**, 60–62 (1990).
21. Maslov, A. V. & Citrin, D. S. Quantum-well optical modulator at terahertz frequencies. *J. Appl. Phys.* **93**, 10131–10133 (2003).
22. Galdin, S., Dollfus, P., Aubry-Fortuna, V., Hesto, P. & Osten, H. J. Band offset predictions for strained group IV alloys: $\text{Si}_{1-x}\text{Ge}_x$ on $\text{Si}(001)$ and $\text{Si}_{1-x}\text{Ge}_x$ on $\text{Si}_{1-x}\text{Ge}_x(001)$. *Semicond. Sci. Technol.* **15**, 565–572 (2000).
23. Rieger, M. M. & Vogl, P. Electronic-band parameters in strained $\text{Si}_{1-x}\text{Ge}_x$ alloys on $\text{Si}_{1-x}\text{Ge}_x$ substrates. *Phys. Rev. B* **48**, 14276–14287 (1993).
24. Schaffler, F. High-mobility Si and Ge structures. *Semicond. Sci. Technol.* **12**, 1515–1549 (1997).
25. Goossen, K. W., Yan, R. H., Cunningham, J. E. & Jan, W. Y. $\text{Al}_x\text{Ga}_{1-x}\text{As}$ - AlAs quantum well surface-normal electroabsorption modulators operating at visible wavelengths. *Appl. Phys. Lett.* **59**, 1829–1831 (1991).
26. Crow, G. C. & Abram, R. A. Monte Carlo simulations of hole transport in SiGe and Ge quantum wells. *Semicond. Sci. Technol.* **15**, 7–14 (2000).
27. Dresselhaus, G., Kip, A. F. & Kittel, C. Cyclotron resonance of electrons and holes in silicon and germanium crystals. *Phys. Rev.* **98**, 368–384 (1955).
28. Lawaetz, P. Valence-band parameters in cubic semiconductors. *Phys. Rev. B* **4**, 3460–3467 (1971).

Acknowledgements We thank V. Lordi for help with photocurrent setup. We also thank J. Fu, T. Krishnamohan and X. Yu for help with device fabrication and material characterization. Finally, we thank G. S. Solomon and D. S. Gardner for discussions. This work was supported by Intel Corporation and the DARPA/ARO EPIC programme.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.-H.K. (yhkuo@stanford.edu).

Unidirectional molecular motor on a gold surface

Richard A. van Delden¹, Matthijs K. J. ter Wiel¹, Michael M. Pollard¹, Javier Vicario¹, Nagatoshi Koumura¹ & Ben L. Feringa¹

Molecules capable of mimicking the function of a wide range of mechanical devices have been fabricated, with motors that can induce mechanical movement attracting particular attention^{1,2}. Such molecular motors convert light or chemical energy into directional rotary or linear motion^{2–10}, and are usually prepared and operated in solution. But if they are to be used as nanomachines that can do useful work, it seems essential to construct systems that can function on a surface, like a recently reported linear artificial muscle¹¹. Surface-mounted rotors have been realized and limited directionality in their motion predicted^{12,13}. Here we demonstrate that a light-driven molecular motor capable of repetitive unidirectional rotation¹⁴ can be mounted on the surface of gold nanoparticles. The motor design¹⁴ uses a chiral helical alkene with an upper half that serves as a propeller and is connected through a carbon–carbon double bond (the rotation axis) to a lower half that serves as a stator. The stator carries two thiol-functionalized ‘legs’, which then bind the entire motor molecule to a gold surface. NMR spectroscopy reveals that two photo-induced *cis-trans* isomerizations of the central double bond, each followed by a thermal helix inversion to prevent reverse rotation, induce a full and unidirectional 360° rotation of the propeller with respect to the surface-mounted lower half of the system.

Inspired by the ATP-ase system¹⁵, we constructed an artificial surface-mounted motor schematically shown in Fig. 1a. The design

of the motor molecule **1** is based on a second-generation¹⁴ light-driven rotary motor **2** with a symmetric lower half bearing two methoxy substituents (Fig. 1b). Replacing these groups by two C₈-spacers terminated with thiols (as shown in structure **1**) allowed self-assembly onto a gold surface, providing **1-Au**. Gold nanoparticles are particularly appropriate for our purpose, as chromophore functionalized nanoparticles are well studied¹⁶ and photochromism of azobenzenes¹⁷ and electrochemical switching of rotaxanes¹⁸ attached to such nanoparticles has been demonstrated. Two points of attachment are essential to prevent uncontrolled thermal rotation of the entire system with respect to the surface. The C₈-spacer should diminish direct (electronic) interaction between the chromophores and the Au surface (which might influence the excited state processes) and give the separate photoactive moieties sufficient free volume to perform the anticipated rotary motion. On the basis of the dynamic processes in structurally related molecular motors¹⁴, **1-Au** was expected to exhibit photochemical and thermal isomerization processes, as shown in Fig. 1c.

Two energetically uphill photochemical isomerization steps (steps 1 and 3 in Fig. 1) each followed by an energetically downhill irreversible thermal helix inversion step (steps 2 and 4 in Fig. 1) result in a full 360° rotation of one half of the molecule with respect to the other. The direction of rotation is controlled by the configuration at the stereogenic centre. Crucial is a strong energetic preference for the methyl substituent to adopt a pseudo-axial orientation. Irradiation

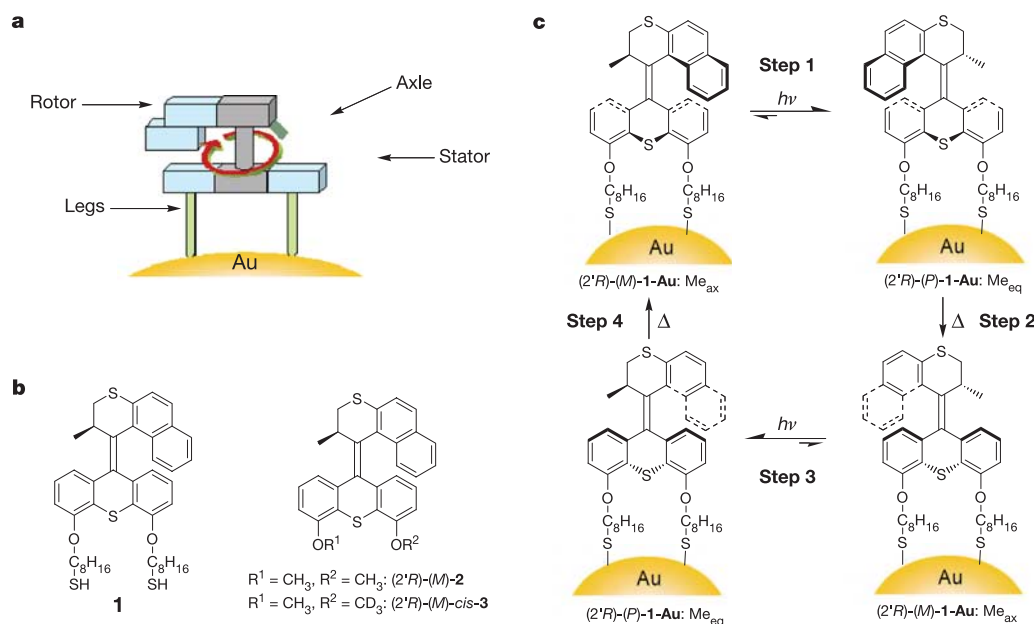


Figure 1 | Molecular motor anchored to a surface. **a**, Design of a surface-bound rotary motor. The system consists of a rotor connected via an axle (axis of rotation) to a stator part that is bound to a gold surface via two legs. **b**, Structure of motor **1** for surface studies and **2**, **3** for solution studies; **1-Au** denotes motor molecule **1** assembled onto Au. *R* denotes absolute configuration at the stereogenic centre; *M* and *P* denote helicity of the molecule. **c**, The four-state unidirectional rotation of functionalized nanoparticle **1-Au** is shown ($h\nu$, photochemical step; Δ , thermal step). The photoisomerizations were induced by irradiation at $\lambda \geq 280$ nm or $\lambda = 365$ nm. Me_{ax} indicates the pseudo-axial orientation of the methyl substituent, Me_{eq} indicates the unstable pseudo-equatorial orientation of the methyl substituent.

¹Department of Organic Chemistry, Stratingh Institute, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands.

of the stable (2'*R*)-(M)-isomer of **1-Au** with a pseudo-axial methyl substituent effects a *cis* → *trans* isomerization, which inverts the helicity of the system to form the (2'*R*)-(P)-isomer. Consequently, the methyl substituent is forced to adopt an energetically disfavoured pseudo-equatorial orientation. Upon heating, this unstable isomer undergoes a helix inversion resulting in the formation of the initial (2'*R*)-(M)-isomer, after a net 180° rotation of the upper half of the molecule in an anticlockwise fashion with respect to the lower half. This sequence is repeated in steps 3 and 4 to complete a full rotary cycle. The unidirectionality of the rotary process is ensured by the irreversibility of the thermal helix inversion steps.

Target compound **2** (4,5-dimethoxy-9-(2',3'-dihydro-2'-methyl-1'*H*-naphtha[2,1-*b*]thiopyran-1'-ylidene)-9*H*-thioxanthene; Fig. 1b), which serves as a precursor and a solution phase model for **1-Au**, was synthesized by adapted methods used for second-generation motors¹⁰. The (2'*R*)-(M)-**2** stereoisomer was resolved by chiral high-performance liquid chromatography (HPLC) and the absolute stereochemistry unequivocally assigned based on circular dichroism (CD) spectroscopy and X-ray analysis (see Supplementary Table 1). Starting from (2'*R*)-(M)-**2**, deprotection of the methoxy-substituents and introduction of two octylthiol moieties was achieved by standard synthetic techniques. Nanoparticles (2'*R*)-(M)-**1-Au** were prepared by the Brust-Schiffrin method¹⁹ and characterized by dynamic light scattering (DLS), transmission electron microscopy (TEM) (Supplementary Fig. 11), and spectroscopic methods (for ultraviolet/visible (UV/Vis) and CD spectroscopy see Fig. 2, and for additional CD, UV, NMR, Fourier transform infrared (FT-IR) and surface enhanced Raman spectroscopy, see Supplementary Figs 2, 3, 4, 7–9 and 10, respectively).

The UV/Vis and CD spectra of (2'*R*)-(M)-**1-Au** and (2'*R*)-(M)-**2**

are characteristic of helical overcrowded alkenes (Fig. 2). The CD spectra of **1-Au** and **2** (Fig. 2a) are nearly identical because they solely reflect the helical chirality of the diaryl alkene moieties and confirm the (M)-helicity of **1-Au** and **2**. The UV/Vis spectrum of **1-Au** (Fig. 2b) is a superposition of the spectral bands of the motor moieties and a broad absorption of the nanoparticle, reaching far into the visible region (see Supplementary Fig. 2).

TEM data, together with the molar quantity of olefin moieties per gram of nanoparticles (as determined by CD spectroscopy) and the known density of gold (19.3 g cm^{-3})²⁰, gives the average overall formula of one nanoparticle of **1-Au**: $\text{Au}_{251}((\text{S}(\text{CH}_2)_8\text{O})_{22}\text{C}_{27}\text{H}_{18}\text{S}_2)_{26}$. It can be calculated that for **1** there is only $\sim 0.23 \text{ nm}^2$ surface area per linked sulphur atom; this can be compared with the surface area of alkanethiols on a flat Au surface as determined by electron diffraction studies (0.214 nm^2)²¹.

The combined photochemical and thermal processes were investigated spectroscopically in toluene solution. The presented data for **1-Au** are in full agreement with those of the parent (model and control) compound **2**, unless indicated.

A $1.035 \times 10^{-5} \text{ M}$ solution (chromophore concentration) of (2'*R*)-(M)-**1-Au** in toluene at room temperature was irradiated (wavelength $\lambda \geq 280 \text{ nm}$) until a photostationary state (PSS) was reached. The approximate inversion of the CD spectrum upon irradiation (Fig. 2a) indicates the change in helicity going from (2'*R*)-(M)-**1-Au** to (2'*R*)-(P)-**1-Au** (step 1 in Fig. 1c). In the UV/Vis spectrum a decrease in the intensity of the high wavelength band is indicative of the formation of (2'*R*)-(P)-**1-Au** (Fig. 2b). Clear isosbestic point(s) in both CD and UV/Vis spectra for the photochemical and thermal steps support clean isomerization processes (Fig. 2). From the UV/Vis spectra of (2'*R*)-(M)-**2** and (2'*R*)-(P)-**2**, the ideal wavelength for the formation of the unstable isomer was determined to be 365 nm. Subsequent irradiation of (2'*R*)-(M)-**1-Au** at this wavelength indeed resulted in a more selective isomerization with an associated increase in intensity of the CD bands (Fig. 2a) and an anticipated PSS of approximately 94:6. UV/Vis and CD data are consistent with this observation (Fig. 2).

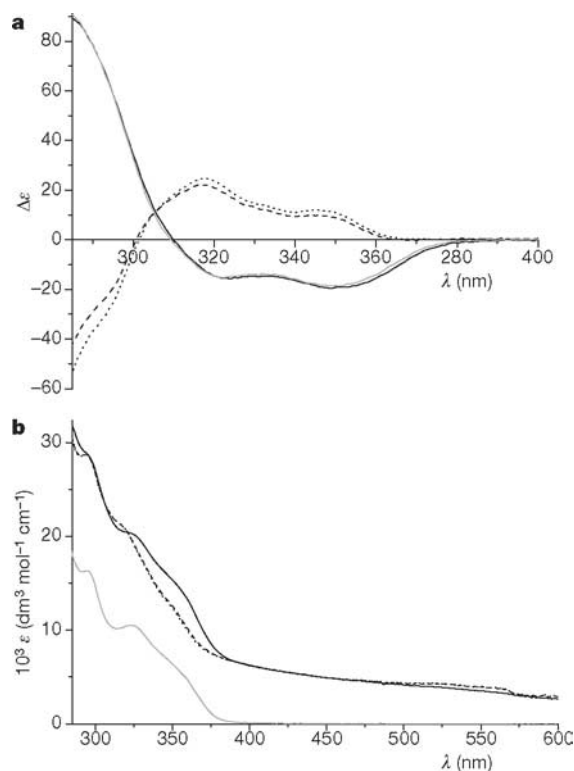


Figure 2 | CD and UV spectra of **1-Au and **2**.** CD (a) and UV/Vis (b) spectra of pure (2'*R*)-(M)-**1-Au** (solid black lines), PSS_{≥280 nm} (dashed black) and PSS_{365 nm} (dotted black) samples (all spectra are adjusted for molar concentration of chromophores), and CD (a) and UV/Vis (b) spectrum of (2'*R*)-(M)-**2** (solid grey) in toluene. After heating of the PSS samples ($T > 50^\circ \text{C}$), the original spectra of (2'*R*)-(M)-**1-Au** and (2'*R*)-(M)-**2** were obtained.

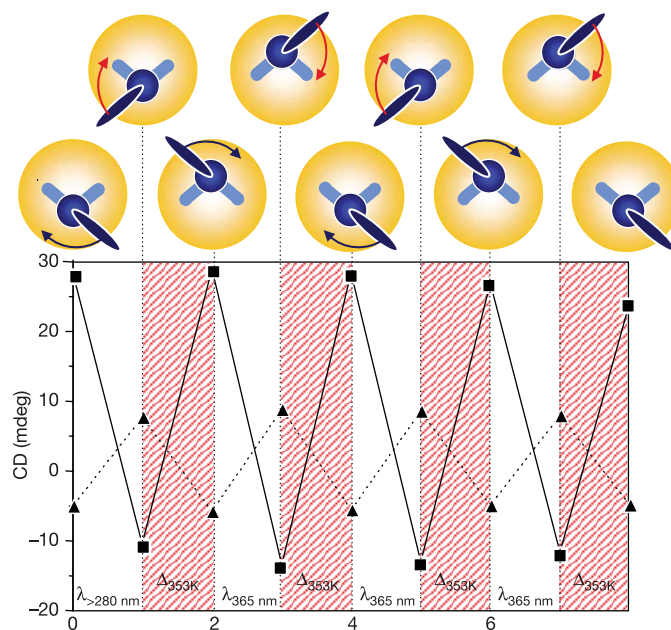


Figure 3 | Following two full turns by CD spectroscopy. Schematic representation of the unidirectional rotation of **1** (as viewed along the rotation axis) and two full four-stage 360° rotary cycles followed by CD spectroscopy. The change in CD intensity (mdeg) at 290 nm (solid line) and 320 nm (dashed) at each photochemical ($h\nu_{\lambda \geq 280 \text{ nm}}$ and $h\nu_{\lambda = 365 \text{ nm}}$) and thermal (Δ_{353K}) isomerization step is shown.

Subsequent heating of the PSS sample of (2'*R*)-(P)-1-Au ($T \geq 50^\circ\text{C}$) resulted in full conversion of unstable (2'*R*)-(P)-1-Au to stable (2'*R*)-(M)-1-Au, as a result of the thermal helix inversion (step 2 in Fig. 1c). It was verified that rotary motion occurs while motor 1-Au remains bound to the surface; this was indicated by the presence of exclusively broad signals in the ^1H NMR spectrum (no trace of free motor in solution could be detected) after either heating the sample at 70°C for 2 h or irradiation for 3 h at 365 nm. Subsequent KCN mediated etching of the gold core of this sample of 1-Au after 3 h irradiation at 365 nm revealed unbound unstable and stable motor in a 2.5:1 ratio.

The kinetics of this helix inversion were determined by monitoring

the change in CD intensity with time at various temperatures ($T = 323, 333, 343, 353\text{ K}$), providing the rate constant (k_t) which was used to calculate the Gibbs free energy of activation ($\Delta^\ddagger G^\circ = 96 \pm 2\text{ kJ mol}^{-1}$) for the (2'*R*)-(P)-1-Au to (2'*R*)-(M)-1-Au interconversion using the Eyring equation. The slightly higher barrier than found for 2 in solution ($\Delta^\ddagger G^\circ = 94 \pm 2\text{ kJ mol}^{-1}$) can be attributed to the reduction in the degrees of freedom of the molecule when grafted onto the Au surface. At room temperature (293 K), the half-life for thermal helix inversion has almost doubled going from 2 ($t_{1/2} = 56 \times 10^2\text{ s}$) to 1-Au ($t_{1/2} = 12 \times 10^3\text{ s}$). After the light-induced energetically uphill process (2'*R*)-(M)-1-Au to (2'*R*)-(P)-1-Au (step 1 in Fig. 1c) followed by a thermal energetically

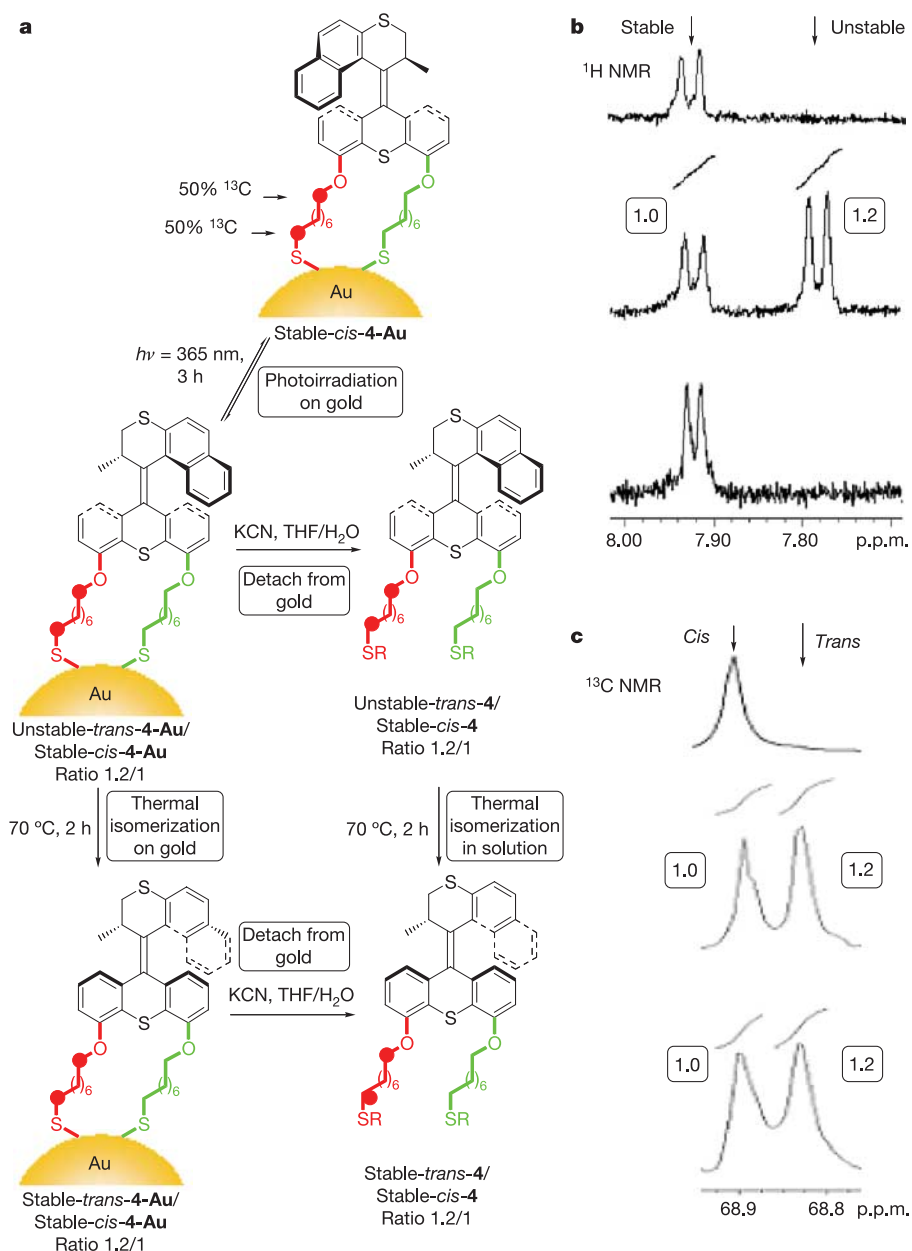


Figure 4 | Confirmation of unidirectional rotation on the nanoparticle using NMR spectroscopy. **a**, Photochemical and thermal isomerization of ^{13}C isotope labelled motor 4 on gold (4-Au) and structures of major isomers of 4-Au. Cleaved motor probably exists as a mixture of free thiol ($\text{R} = \text{H}$) and disulphide ($\text{R} = \text{S-Motor}$). **b**, Key signals (ArH) in the ^1H NMR of stable-*cis*-4 in solution (top panel), after irradiation of stable-*cis*-4-Au and core etching (middle panel) and after irradiation of stable-*cis*-4-Au, heating, then etching (bottom panel). **c**, Key signals ($-\text{O}^{13}\text{CH}_2[\text{CH}_2]_7$) in the ^{13}C

NMR of stable-*cis*-4 in solution (top panel), after irradiation of stable-*cis*-4-Au and core etching, and subsequent heating (middle panel), and after irradiation of stable-*cis*-4-Au, heating, then etching (bottom panel). The conversion of stable isomer 4-Au to unstable isomer 4-Au (evaluated by ^1H NMR) matched the final conversion of stable-*cis*-4-Au to stable-*trans*-4-Au evaluated by ^{13}C NMR. In both cases, NMR analysis was performed after detachment.

downhill helix inversion ($2'R$)-(P)-**1-Au** to ($2'R$)-(M)-**1-Au** (step 2 in Fig. 1c), the upper rotor half of the molecule has performed a half rotation in an anticlockwise direction with respect to the lower half. Performing the same two-step process repeatedly results in continual unidirectional rotary motion. The state of the motor can be followed by CD spectroscopy, as is shown (Fig. 3) for two consecutive 360° cycles.

To confirm that the introduction of the two legs in the lower part of the molecule does not affect the four-stage unidirectional rotary process, we studied ($2'R$)-(M)-**3**, which has a non-symmetrical lower part with one OCH₃ and one OCD₃ (see Fig. 1b). These groups have negligible influence on the photochemical and thermal behaviour, and allowed the detection by NMR of all four steps in the rotary cycle (see Supplementary Information, Section 3).

Unequivocal proof for the unidirectionality of the motor attached to gold nanoparticles was achieved with isotope labelled motor **4-Au** (Fig. 4). The only difference between **1-Au** and **4-Au** is that one leg in the latter compound contains a ¹³C isotope at the two positions flanking the stator and the surface bound thiol. The two distinct legs created in this way allow accurate detection by ¹H and ¹³C NMR of the isomers after photochemical and thermal steps on the surface, as illustrated in Fig. 4a.

The stable *cis*-isomer of **4** bound to the gold (*cis*-**4-Au**) was irradiated at 365 nm wavelength, and the sample split into two portions. The first portion was treated (KCN/THF/H₂O) so as to detach the motor from the surface to provide a mixture of the *trans*-unstable-**4** and *cis*-stable-**4** (ratio 1.2:1, ¹H NMR analysis, Fig. 4b). Subsequent thermal isomerization in toluene solution resulted in stable *trans*-**4** and stable-*cis*-**4** isomers (1.2:1 ratio, ¹³C NMR analysis Fig. 4c).

The second portion was heated (70 °C, 2 h) to induce thermal isomerization of the motor while still on the gold surface (unstable-*trans*-**4-Au** to stable-*trans*-**4-Au**). Subsequent cleavage (KCN/THF/H₂O) of the motor from the gold provided an identical mixture of stable-*trans*-**4** and stable-*cis*-**4**. These data show that the photochemical conversion of the stable form of **4-Au** to the unstable form of **4-Au** (¹H NMR analysis) correlates perfectly with the formation of stable *trans*-**4-Au** from stable *cis*-**4-Au** (¹³C NMR analysis).

The combined results (from CD studies of optically active ($2'R$)-(M)-**1-Au**, and NMR studies of ¹³C isotopically labelled **4-Au**, with distinct legs) provide compelling evidence for the unidirectionality of the rotary process on the surface. We expect that the successful formation of fully functional surface-mounted rotors will enable investigation of the concerted action of a large ensemble of unidirectional molecular motors, and that this system might be a first step towards the construction of more elaborate and functional nanosized mechanical devices.

Received 12 July; accepted 2 August 2005.

1. Balzani, V., Venturi, M. & Credi, A. *Molecular Devices and Machines—A Journey into the Nanoworld* (Wiley-VCH, Weinheim, 2003).

2. van Delden, R. A., ter Wiel, M. K. J., Koumura, N. & Feringa, B. L. in *Molecular Motors* (ed. Schliwa, M.) Ch. 23, 559–577 (Wiley-VCH, Weinheim, 2003).
3. Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. Light-driven monodirectional molecular rotor. *Nature* **401**, 152–155 (1999).
4. Kelly, T. R., De Silva, H. & Silva, R. A. Unidirectional rotary motion in a molecular system. *Nature* **401**, 150–152 (1999).
5. Leigh, D. A., Wong, J. K. Y., Dehez, F. & Zerbetto, F. Unidirectional rotation in a mechanically interlocked molecular rotor. *Nature* **424**, 174–179 (2003).
6. Anelli, P. L., Spencer, N. & Stoddart, J. F. A molecular shuttle. *J. Am. Chem. Soc.* **113**, 5131–5133 (1991).
7. Sherman, W. B. & Seeman, N. C. A precisely controlled DNA biped walking device. *Nano Lett.* **4**, 1203–1207 (2004).
8. Sauvage, J.-P. (ed.) *Molecular Machines and Motors* (Structure and Bonding, Vol. 99, Springer, Berlin, 2001).
9. van Delden, R. A., Koumura, N., Harada, N. & Feringa, B. L. Unidirectional rotary motion in a liquid crystalline environment; colour tuning by a molecular motor. *Proc. Natl Acad. Sci. USA* **99**, 4945–4949 (2002).
10. Koumura, N., Geertsema, E. M., van Gelder, M. B., Meetsma, A. & Feringa, B. L. Second generation light-driven molecular motors. Unidirectional rotation controlled by a single stereogenic center with near-perfect photoequilibrium and acceleration of the speed of rotation by structural modification. *J. Am. Chem. Soc.* **124**, 5037–5051 (2002).
11. Liu, Y. et al. Linear artificial muscles. *J. Am. Chem. Soc.* **127**, 9745–9759 (2005).
12. Zheng, X. et al. Dipolar and nonpolar altitudinal molecular rotors mounted on an Au(111) surface. *J. Am. Chem. Soc.* **126**, 4540–4542 (2004).
13. Kottas, G. S., Clarke, L. I. & Horinek, D. Michl, J. Artificial molecular rotors. *Chem. Rev.* **105**, 1281–1376 (2005).
14. Koumura, N., Geertsema, E. M., Meetsma, A. & Feringa, B. L. Light-driven molecular rotor: unidirectional rotation controlled by a single stereogenic center. *J. Am. Chem. Soc.* **122**, 12005–12006 (2000).
15. Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. Direct observation of the rotation of F₁-ATPase. *Nature* **386**, 299–302 (1997).
16. George Thomas, K. & Kamat, P. V. Chromophore-functionalised gold nanoparticles. *Acc. Chem. Res.* **36**, 888–898 (2003).
17. Manna, A. et al. Optimised photoisomerisation on gold nanoparticles capped by unsymmetrical azobenzene disulfides. *Chem. Mater.* **15**, 20–28 (2003).
18. Long, B., Nikitin, K. & Fitzmaurice, D. Assembly of an electronically switchable rotaxane on the surface of a titanium dioxide nanoparticle. *J. Am. Chem. Soc.* **125**, 15490–15498 (2003).
19. Brust, M., Walker, M., Behell, D., Schiffrin, D. J. & Whyman, R. J. Synthesis of thiol-derivatised gold nanoparticles in a 2-phase liquid-liquid system. *Chem. Soc. Chem. Comm.*, 801–802 (1994).
20. Lide, D. R. (ed.) *CRC Handbook of Chemistry and Physics* 2003–2004 84th edn 4–59 (CRC Press, Boca Raton, 2003).
21. Ullman, A. Formation and structure of self-assembled monolayers. *Chem. Rev.* **96**, 1533–1554 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. R. van den Brom for performing the TEM measurements described, A. Meetsma for X-ray analysis and J. G. McGarvey for access to Raman facilities. Financial support from the Netherlands Organisation for Scientific Research (NWO-CW), the Materials Science Centre, and the University of Groningen is acknowledged. J.V. thanks the Departamento de Educación, Universidades e Investigación del Gobierno Vasco, for a postdoctoral fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.L.F. (B.L.Feringa@rug.nl).

Low-latitude seasonality of Cretaceous temperatures in warm and cold episodes

Thomas Steuber¹, Markus Rauch¹, Jean-Pierre Masse², Joris Graaf³ & Matthias Malkoč¹

The Cretaceous period is generally considered to have been a time of warm climate^{1–6}. Evidence for cooler episodes exists, particularly in the early Cretaceous period^{6–8}, but the timing and significance of these cool episodes are not well constrained. The seasonality of temperatures is important for constraining equator-to-pole temperature gradients and may indicate the presence of polar ice sheets; however, reconstructions of Cretaceous sea surface temperatures are predominantly based on the oxygen isotopic composition of planktonic foraminifera^{1–4} that do not provide information about such intra-annual variations. Here we present intra-shell variations in $\delta^{18}\text{O}$ values of rudist bivalves (Hippuritoidea) from palaeolatitudes between 8° and 31° N, which record the evolution of the seasonality of Cretaceous sea surface temperatures in detail. We find high maximum temperatures (~35 to 37 °C) and relatively low seasonal variability (<12 °C) between 20° and 30° N during the warmer Cretaceous episodes. In contrast, during the cooler episodes our data show seasonal sea surface temperature variability of up to 18 °C near 25° N, comparable to the range found today. Such a large seasonal variability is compatible with the existence of polar ice sheets.

The Cretaceous was a period of warm climate, with mid-Cretaceous (100–88 million years, Myr, ago) polar sea surface temperature (SST) as high as 20 °C (ref. 9). Although there is evidence for polar ice in the early Cretaceous^{6–8}, greenhouse conditions of Earth's climate system during the middle and late Cretaceous are supported by a plethora of geological, palaeontological and geochemical data⁶. Geochemical proxies, including $\delta^{18}\text{O}$ values of foraminifera, fish teeth and the composition of membrane lipids^{1–4,9–11}, are from biological materials that do not provide data on sub-annual timescales. Reconstructed Cretaceous SSTs are thus believed to represent annual averages, although the seasonal timing of formation of the studied biological materials is not known. Estimates of terrestrial seasonality based on fossil floras are an exception, but are only available for a few localities¹².

We have analysed intra-shell variations in isotopic ($\delta^{18}\text{O}$, $\delta^{13}\text{C}$) and chemical (Mg, Sr, Fe, Mn) compositions of shells of rudist bivalves from localities (Fig. 1) in the Caribbean, the peri-Mediterranean region and the Middle East (8–31° N palaeolatitude, 85° W–31° E palaeolongitude). Numerical ages for most studied localities were obtained by strontium isotope stratigraphy¹³, the others have a precise biostratigraphical control. Details of localities, stratigraphy and taxonomy of the studied shells are available as Supplementary Information. Most of our data are from 15–30° N palaeolatitude (Fig. 1), and thus from regions that must have experienced the highest seasonal temperature extremes. Rudist bivalves inhabited marine subtidal environments well above the thermocline¹⁴. In contrast to foraminifera, which have been the main source of proxy data on Cretaceous SST, the outer shell layer of certain groups of rudist is compact (see Supplementary Information), and has a

large potential to preserve the original chemical and isotopic composition^{15,16}. Modern bivalve molluscs reliably record temperature-dependent fractionation of oxygen isotopes in the $\delta^{18}\text{O}$ values of their shell carbonates^{17,18}. Metabolic effects can thus be largely ruled out for rudist bivalves, although pronounced disequilibrium fractionation has been reported for a single genus¹⁵.

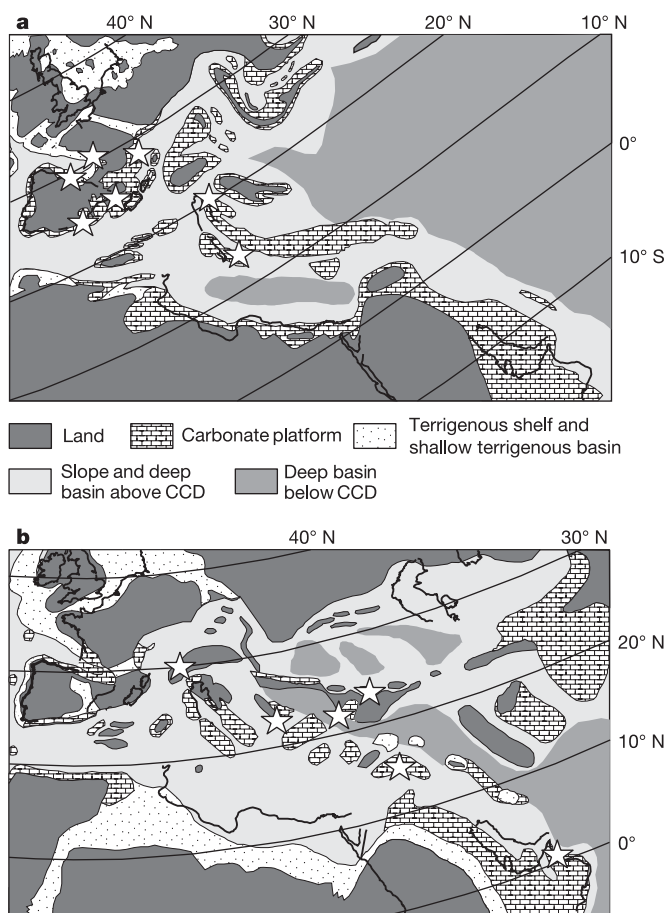


Figure 1 | Palaeo-positions of localities studied. White stars indicate localities; those in the Caribbean (Jamaica) are not shown. Modern coastlines—that is, of Great Britain in the upper left, and the Arabian Gulf in the lower right, of the maps—are given for orientation. **a**, Localities from 120–91 Myr ago plotted on early Aptian (119 Myr ago) palaeogeography, modified from ref. 25. **b**, Localities from 90–66 Myr ago plotted on Maastrichtian (67 Myr ago) palaeogeography, modified from ref. 26. See Supplementary Information for precise palaeolatitudes.

¹Institut für Geologie, Mineralogie und Geophysik, Ruhr-Universität, 44801 Bochum, Germany. ²Centre de Sédimentologie-Paléontologie, Université de Provence, 13331 Marseille Cedex 03, France. ³Department of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands.

All studied shells show relatively small variations in $\delta^{13}\text{C}$ values and Sr concentrations, with $\delta^{13}\text{C}$ values within the range of normal marine carbonates. Samples with Mn and Fe concentrations above 100 p.p.m. have been excluded from further evaluation, because high Mn and Fe are considered to indicate diagenetic alteration¹⁵. In most specimens, all samples had less than 20 p.p.m. Mn and Fe. All shells yielded cyclic variations of $\delta^{18}\text{O}$ values (Fig. 2), except for those from low palaeolatitudes which have small ranges of $\delta^{18}\text{O}$ values with no distinct pattern. The cyclic variations in $\delta^{18}\text{O}$ values follow a sine function in most specimens (Fig. 2), which traces the seasonal variation of incident solar radiation at non-equatorial latitudes. This broadly symmetrical sinusoidal pattern implies rather constant growth rates and environmental conditions, which, although varying regularly, were favourable throughout the year¹⁵. Annual cycles are covered by up to 55 samples, so that one sample represents an average of ambient temperatures during one week of shell growth. Several of the specimens studied yielded incomplete records of intra-annual isotopic variation, either because seasonal growth cessation, indicated by prominent growth bands, produced truncated seasonality minima or maxima, or because the available fragments of shells recorded sub-annual growth increments only (Fig. 3).

Our data indicate a distinct drop in mean SSTs during the Aptian, with seasonal maxima decreasing by 5 °C (Fig. 3). Specimens from 17° N palaeolatitude have similar mean values but a smaller intra-annual range than shells from 25° N, which have recorded the largest seasonal temperature variations (of up to 18 °C) among the studied shells (Fig. 4). Lowest temperatures are derived from late Aptian shells. Data coverage is relatively poor during the Albian–Cenomanian, but there is a trend of increasing temperatures, and four specimens from the North Atlantic (western France and northern Spain) have recorded lower SST than Tethyan representatives (Fig. 3). Shells from the upper Turonian–Coniacian and lower Campanian indicate high seasonal extremes that range up to 37 °C, and a low intra-annual variation with minimum temperatures of 23–25 °C (Figs 3, 4). Significantly lower SSTs recorded in a single mid-Coniacian specimen correspond to a pronounced drop in temperature reported from southern intermediate latitudes¹⁹. Climatic cooling to conditions similar to those of the late Aptian–early Albian is seen in a single specimen from the mid-Campanian (27° N palaeolatitude). The few studied shells from the mid-Campanian–Maastrichtian are from low palaeolatitudes (8–14° N), show no distinct cycles of $\delta^{18}\text{O}$ values and Mg concentrations, and have recorded moderate intra-annual temperature variations of 27–35 °C (Fig. 4). Changes in mean

temperature as derived from shells, which apparently recorded seasonal minima and maxima of SST, agree well with the most continuous Cretaceous temperature record available, which is from southern intermediate palaeolatitude¹⁹, and also with the broad trends derived from phosphatic fish remains of intermediate northern palaeolatitudes¹⁰. Mean temperatures do not exceed 32 °C, and are between 20 and 25 °C for most of the early Cretaceous, which is consistent with previous reports of mean tropical SST derived from skeletal carbonates^{1–4,20}.

Potential errors of $\delta^{18}\text{O}$ -based SST estimates²¹ are related predominantly to uncertainty about the $\delta^{18}\text{O}$ value of ancient sea water ($\delta^{18}\text{O}_{\text{sw}}$) due to latitudinal differences in the precipitation–evaporation balance. Model simulations for the mid-Cretaceous suggest a steep gradient from -1.1‰ $\delta^{18}\text{O}$ (VSMOW) at 15° N to -0.5‰ $\delta^{18}\text{O}$ (VSMOW) at 30° N (ref. 22). Adopting this gradient increases $\delta^{18}\text{O}$ -derived SSTs by 2 °C at 30° N (Fig. 4). Seasonal variations in the $\delta^{18}\text{O}_{\text{sw}}$ can have a larger effect on the calculated SSTs. Generally, the intra-annual temperature range would increase in case of reduced salinity during the warm season, and elevated salinity during the cool season. Owing to Cretaceous palaeogeography, the intertropical convergence zone was shifted southwards, and the subtropical arid belt ran along the southern margin of Tethys²³. Most of our data are from the northern Tethyan margin so that an inverse seasonal covariance of SST and salinity is unlikely. Conversely, the $\delta^{18}\text{O}$ -derived intra-annual temperature variation would be smaller in case of a positive seasonal salinity–SST covariance. Although this is a more likely assumption, corrections based on intra-annual salinity variations are highly speculative. Suffice it to say that palaeotemperatures at seasonal extremes would change by 3.5 °C were salinity to have differed from normal marine conditions by 5 p.s.u. (ref. 15). This figure is considered the maximum deviation from normal marine conditions in view of the associated biota, such as calcareous algae, benthic foraminifera, and colonial corals at most of the studied localities. The $\delta^{13}\text{C}$ values of rudist calcite can further constrain the magnitude of intra-annual salinity changes. Whereas a salinity increase of 5 p.s.u. would not change the $\delta^{13}\text{C}$ value of dissolved inorganic carbon ($\delta^{13}\text{C}_{\text{DIC}}$), a salinity reduction of 5 p.s.u. caused by freshwater influx would lower the $\delta^{13}\text{C}_{\text{DIC}}$ by $\sim 2.5\text{‰}$. Shells with covariant curves of $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ values have much lower ranges in $\delta^{13}\text{C}$ (Fig. 2b), and these variations could also result from biological processes affecting the local marine DIC. A significant over-estimation of SST because of seasonally reduced salinity can thus be ruled out, whereas the effects of seasonally

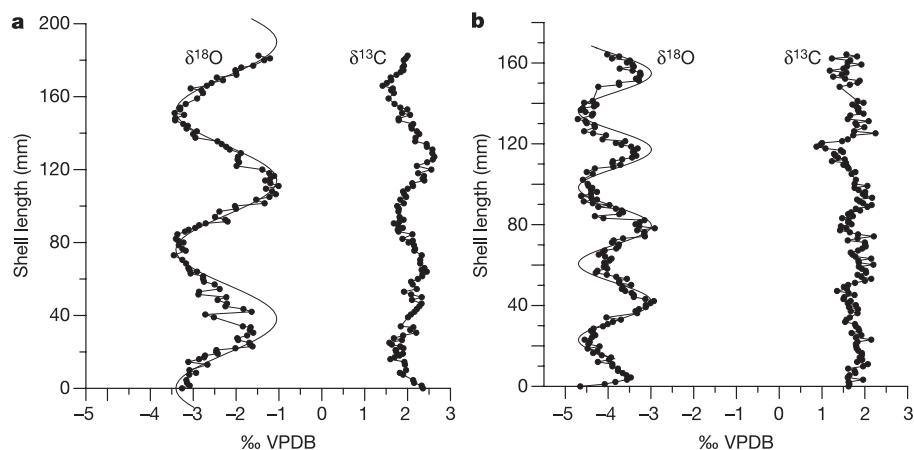


Figure 2 | Intra-shell variation of $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ values in sclerochronological sections. **a**, Early Aptian cool period, maximum seasonality. Palaeolatitude, 25.5° N. *Toucasia carinata* (Matheron), lower Aptian of Mts de Vaucluse (southern France), specimen Ma42. **b**, Late Campanian warm period, minimum seasonality. Palaeolatitude, 25.9° N. *Vaccinites alpinus* (Douville), lower Campanian of Pontid Mts (northern

Turkey), specimen H587. Sine function fitted to curves of $\delta^{18}\text{O}$ values indicates continuous, almost constant, rates of shell accretion and a complete record of intra-annual temperature–salinity variations.

$$\delta^{18}\text{O} = \left\{ \left[\frac{(^{18}\text{O}/^{16}\text{O})_{\text{sample}}}{(^{18}\text{O}/^{16}\text{O})_{\text{standard}}} - 1 \right] \times 1,000 \right\};$$

$$\delta^{13}\text{C} = \left\{ \left[\frac{(^{13}\text{C}/^{12}\text{C})_{\text{sample}}}{(^{13}\text{C}/^{12}\text{C})_{\text{standard}}} - 1 \right] \times 1,000 \right\}.$$

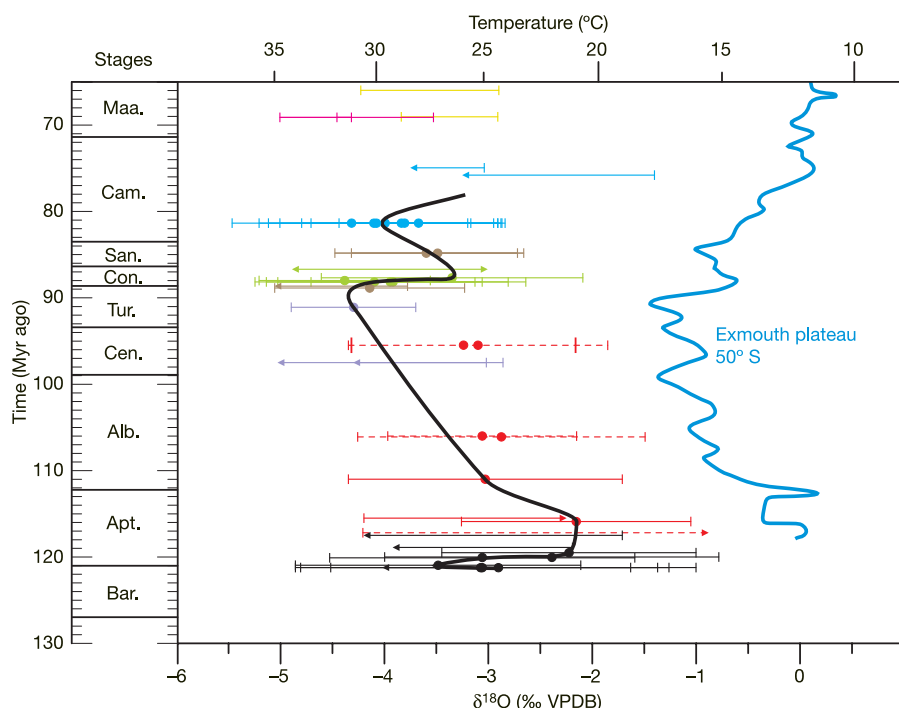


Figure 3 | Ranges of $\delta^{18}\text{O}$ values in sclerochronological sections of rudist bivalves. Samples are from Jamaica (yellow), Oman (pink), Turkey (blue), Austria (brown), central Greece (green), Croatia and Ionian Islands, Greece (violet), southern Spain (red), northern Spain and western France (dashed red lines), and southern France (black). See Supplementary Information for details of localities, stratigraphy and geochemical data. Arrows indicate incomplete records of annual minima and/or maxima of $\delta^{18}\text{O}$ values. Mean values are indicated by circles for those shells that most probably recorded complete intra-annual variation. Broken red lines denote localities

(northern Spain, western France) that were not considered in the curve (black) of the evolution of mean annual SST of Tethys, because they represent cooler water masses of the northern Atlantic Ocean (Fig. 1). Blue curve to the right is the evolution of the $\delta^{18}\text{O}$ values of the fine fraction of calcareous ooze at $\sim 50^\circ\text{S}$ palaeolatitude (Exmouth plateau), corrected for diagenetic effects and assumed to represent mean annual SSTs (ref. 19). All $\delta^{18}\text{O}$ values are converted to palaeotemperatures after ref. 27, assuming -1‰ $\delta^{18}\text{O}$ (VSMOW) of Cretaceous sea water. Timescales are of ref. 28 (130–98.5 Myr ago) and ref. 29 (98.5–65 Myr ago).

increased $\delta^{18}\text{O}_{\text{sw}}$ cannot be constrained and could cause an underestimation of annual SST maxima of up to 3.5°C .

The high seasonality of up to 18°C recorded in the shells of Barremian–Aptian rudist bivalves is compatible with the existence of polar ice sheets, as evidenced by glacial deposits in high latitudes^{7,8}. Reconstructed SSTs for this episode (100–130 Myr ago) are even lower than today, when the $\delta^{18}\text{O}_{\text{sw}}$ is adjusted to ice-free conditions (Fig. 4). They fall into the range of modern seasonality when

-0.66‰ $\delta^{18}\text{O}_{\text{sw}}$ (VSMOW) is assumed, corresponding to one-third the extent of the modern polar ice sheets, an amount estimated from the distribution of glacial sediments for this time⁸. The observation that seasonal temperature extremes at $17\text{--}25^\circ\text{N}$ during the late Barremian–mid-Albian cool period are similar to today is difficult to explain without considering the presence of polar ice. In contrast, the maximum $\delta^{18}\text{O}$ -derived SSTs are higher than the modern extremes for the Cenomanian–Campanian, when ice-free

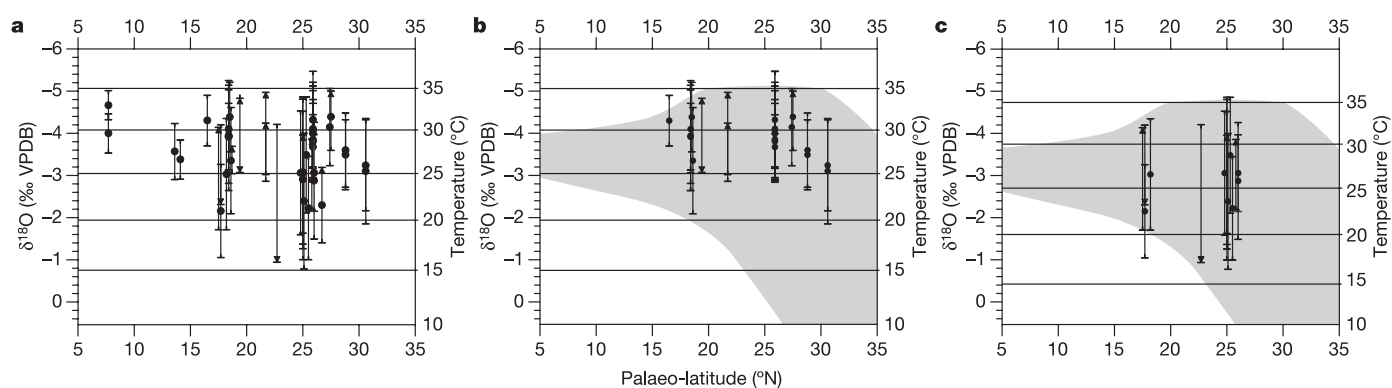


Figure 4 | Intra-shell variations in $\delta^{18}\text{O}$ values and palaeotemperature seasonality as functions of palaeolatitude. Bars indicate ranges between annual minima and maxima of $\delta^{18}\text{O}$ values, and arrows indicate incomplete records of annual minima and/or maxima. Mean values are indicated by circles for shells, which probably recorded complete intra-annual variation. Shaded fields indicate ranges of coastal air temperatures observed today³⁰.

a, Complete data set. **b**, Shells from Cretaceous greenhouse episode (late Albian–early Campanian). **c**, Data from cooler episodes of the Cretaceous (late Barremian–middle Albian). Horizontal lines indicate palaeotemperatures assuming a $\delta^{18}\text{O}_{\text{sw}}$ value of -1‰ VSMOW (**a**, **b**) and -0.66‰ VSMOW (**c**).

conditions are assumed for the $\delta^{18}\text{O}_{\text{sw}}$ value (Fig. 4). A much lower seasonality during the Cretaceous warm period at 25–30°N than today (Fig. 4) is compatible with a low equator-to-pole thermal gradient, which is well documented for the mid-Cretaceous greenhouse^{6,9}. Our data are scarce for the warmest episodes of the mid-Cretaceous, when mean SST of the Arctic Ocean is estimated to have ranged up to 20 °C (ref. 9), and tropical SST may have reached 34 °C (ref. 3).

SST seasonality derived from rudist shells is generally larger than that predicted by climate models^{22,23}. The reliability of climate models deteriorates with increasing latitude²⁴, so that proxy data on SST seasonality at subtropical palaeolatitudes provide important constraints on the dynamics of heat transfer from low to high latitudes, and may help to resolve the discrepancies between latitudinal temperature gradients derived from proxy data and climate models; the existence of such discrepancies is one of the major problems in the understanding of the controlling factors behind Cretaceous climate⁵.

METHODS

Sampling, elemental and isotopic compositions. Sample material of rudist shells was collected from polished thick sections (~2 mm). The outer shell layer (see Supplementary Fig. 1) was sampled in sclerochronological profiles, that is, parallel to the growth axis, with a Merchantek MicroMill or a hand held dental drill using tungsten drill bits (0.5–0.7 mm diameter). Altered parts of the shell that were cracked or spotted with borings of bioeroders were avoided. Stable carbon and oxygen analyses of sample powders by isotope ratio mass spectrometry had an external precision better than $\pm 0.10\text{‰}$ $\delta^{13}\text{C}$ and $\pm 0.21\text{‰}$ $\delta^{18}\text{O}$ (1 s.d., $n = 1,225$). For most specimens, a split of each sample (~0.75 mg) was carefully weighed and used for analysis of elemental concentrations by inductively coupled plasma-atomic emission spectrometry (ICP-AES). Detection limits of elemental concentrations of samples dissolved in 1 ml HCl (1.25 N) and diluted with 5 ml of distilled water were $1.5 \mu\text{g g}^{-1}$ Mg, $0.7 \mu\text{g g}^{-1}$ Sr, $18 \mu\text{g g}^{-1}$ Fe and $30 \mu\text{g g}^{-1}$ Mn. The external precision of analyses was better than ± 8 relative per cent for Mg, Sr, Fe and Mn (1 s.d., $n = 172$).

Numerical ages. The stratigraphy of all post-Cenomanian shells and of most early Cretaceous samples was derived from strontium-isotope stratigraphy. Strontium was separated by standard ion exchange methods from 1 mg of sample powder drilled from the outer shell layer of rudist bivalves. Strontium-isotope ratios were determined with a Finnigan MAT 262 multicollector mass spectrometer, and adjusted to a ratio of 0.709175 of the USGS EN-1 standard to derive numerical ages from the 'look-up' table of ref. 13. Means of $^{87}\text{Sr}/^{86}\text{Sr}$ values obtained from one or more specimens from each of the studied localities were calculated, and the statistical uncertainty of this mean value (2 s.e.m.) was combined with the uncertainty related to the seawater strontium-isotope curve¹³.

Received 17 March; accepted 27 July 2005.

- Pearson, P. N. *et al.* Warm tropical sea surface temperatures in the Late Cretaceous and Eocene epochs. *Nature* **413**, 481–487 (2001).
- Wilson, P. A. & Norris, R. D. Warm tropical ocean surface and global anoxia during the mid-Cretaceous period. *Nature* **412**, 425–429 (2001).
- Norris, R. D., Bice, K. L., Magno, E. A. & Wilson, P. A. Jiggling the tropical thermostat in the Cretaceous hothouse. *Geology* **30**, 299–302 (2002).
- Huber, B. T., Norris, R. D. & MacLeod, K. G. Deep-sea paleotemperature record of extreme warmth during the Cretaceous. *Geology* **30**, 123–126 (2002).
- Poulsen, C. J. A balmy Arctic. *Nature* **432**, 814–815 (2004).
- Frakes, L. A., Francis, J. E. & Syktus, J. I. *Climate Modes of the Phanerozoic* (Cambridge Univ. Press, Cambridge, 1994).
- Kemper, E. Das Klima der Kreidezeit. *Geol. Jb.* **A 96**, 5–185 (1987).
- Price, G. D. The evidence and implications of polar ice during the Mesozoic. *Earth-Sci. Rev.* **48**, 183–210 (1999).
- Jenkyns, H. C., Forster, A., Schouten, S. & Sinninghe Damsté, J. S. High temperatures in the Late Cretaceous Arctic Ocean. *Nature* **432**, 888–892 (2004).

- Pucéat, E. *et al.* Thermal evolution of Cretaceous Tethyan marine waters inferred from oxygen isotope composition of fish tooth enamels. *Paleoceanography* **18**, 1029, doi:10.1029/2002PA000823 (2003).
- Schouten, S. *et al.* Extremely high sea-surface temperatures at low latitudes during the middle Cretaceous as revealed by archaeal membrane lipids. *Geology* **31**, 1069–1072 (2003).
- Spicer, R. A. in *The Cretaceous World* (ed. Skelton, P. W.) (Cambridge Univ. Press, Cambridge, 2003).
- McArthur, J. M., Howarth, R. J. & Bailey, T. R. Strontium isotope stratigraphy: Lowess Version 3: Best-fit to the marine Sr-isotope curve for 0 to 509 Ma and accompanying look-up table for deriving numerical age. *J. Geol.*, 155–170 (2001).
- Gili, E., Masse, J.-P. & Skelton, P. W. Rudists as gregarious sediment-dwellers, not reef-builders, on Cretaceous carbonate platforms. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **118**, 245–267 (1995).
- Steuber, T. Isotopic and chemical intra-shell variations in low-Mg calcite of rudist bivalves (Mollusca: Hippuritacea)—disequilibrium fractionations and Late Cretaceous seasonality. *Int. J. Earth Sci.* **88**, 551–570 (1999).
- Steuber, T. & Rauch, M. Evolution of the Mg/Ca ratio of Cretaceous seawater—Implications from the composition of biological low-Mg calcite. *Mar. Geol.* **217**, 199–213 (2005).
- Wefer, G. & Berger, W. H. Isotope paleontology: growth and composition of extant calcareous species. *Mar. Geol.* **100**, 207–248 (1991).
- Lécuyer, C., Reynard, B. & Martineau, F. Stable isotope fractionation between mollusc shells and marine waters from Martinique Island. *Chem. Geol.* **213**, 293–305 (2004).
- Clarke, L. J. & Jenkyns, H. C. New oxygen isotope evidence for long-term Cretaceous climatic change in the southern hemisphere. *Geology* **27**, 702 (1999).
- Wilson, P. A. & Opdyke, B. N. Equatorial sea-surface temperatures for the Maastrichtian revealed through remarkable preservation of metastable carbonate. *Geology* **24**, 555–558 (1996).
- Crowley, T. J. & Zachos, J. C. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Wing, S. L.) 50–76 (Cambridge Univ. Press, Cambridge, 2000).
- Poulsen, C. J., Barron, E. J., Peterson, W. H. & Wilson, P. A. A reinterpretation of mid-Cretaceous shallow marine temperatures through model-data comparison. *Paleoceanography* **14**, 679–697 (1999).
- Poulsen, C. J., Barron, E. J., Johnson, C. C. & Fawcett, P. in *Evolution of the Cretaceous Ocean/Climate System* (eds Barrera, E. & Johnson, C. C.) 73–89 (GSA Spec. Pap. 332, Geological Society of America, Boulder, 1999).
- DeConto, R. M., Thompson, S. L. & Pollard, D. in *Warm Climates in Earth History* 21–49 (Cambridge Univ. Press, Cambridge, 2000).
- Masse, J. P. *et al.* in *Atlas Tethys Palaeoenvironmental Maps* (eds Dercourt, J., Ricou, L. E. & Vrielynck, B.) (BEICIP-FRANLAB, Rueil-Malmaison, 1993).
- Camoin, G. *et al.* in *Atlas Tethys Palaeoenvironmental Maps* (eds Dercourt, J., Ricou, L. E. & Vrielynck, B.) (BEICIP-FRANLAB, Rueil-Malmaison, 1993).
- Anderson, T. F. & Arthur, M. A. Stable isotopes of oxygen and carbon and their application to sedimentologic and paleoenvironmental problems. *Soc. Econ. Paleont. Mineral. Short Course* **10**, 1–1–151 (1983).
- Gradstein, F. M. *et al.* A Mesozoic time scale. *J. Geophys. Res.* **B 99**, 24051–24074 (1994).
- Obradovich, J. D. A Cretaceous time scale. *Geol. Assoc. Can. Spec. Pap.* **39**, 379–396 (1993).
- Ivany, L. C. *et al.* Intra-annual isotopic variation in *Venericardia* bivalves: Implications for early Eocene temperature, seasonality, and salinity on the U.S. Gulf Coast. *J. Sedim. Res.* **74**, 7–19 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft. We thank D. Buhl, U. Schulte, B. Gehnen and B. Raczek for advice and support in the laboratory. We also thank J. Mutterlose for discussions. S. Özer and O. F. Geyer provided some of the studied specimens from Turkey and Spain.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.S. (thomas.steuber@ruhr-uni-bochum.de).

Cooling of the Earth and core formation after the giant impact

Bernard J. Wood^{1†} & Alex N. Halliday²

Kelvin calculated the age of the Earth to be about 24 million years by assuming conductive cooling from being fully molten to its current state¹. Although simplistic², his result is interesting in the context of the dramatic cooling that took place after the putative Moon-forming giant impact, which contributed the final ~10 per cent of the Earth's mass^{3,4}. The rate of accretion and core segregation on Earth as deduced from the U–Pb system⁵ is much slower than that obtained from Hf–W systematics^{6–8}, and implies substantial accretion after the Moon-forming impact, which occurred 45 ± 5 Myr after the beginning of the Solar System. Here we propose an explanation for the two timescales^{5,9}. We suggest that the Hf–W timescale reflects the principal phase of core-formation before the giant impact. Crystallization of silicate perovskite in the lower mantle during this phase produced Fe³⁺, which was released during the giant impact¹⁰, and this oxidation resulted in late segregation of sulphur-rich metal into which Pb dissolved readily, setting the younger U–Pb age of the Earth. Separation of the latter metal then occurred 30 ± 10 Myr after the Moon-forming impact. Over this time span, in surprising agreement with Kelvin's result, the Earth cooled by about 4,000 K in returning from a fully molten to a partially crystalline state.

The rates and timing of the processes by which the Earth accreted and differentiated are generally studied using chronometers such as ¹²⁹I–¹²⁹Xe, ¹⁸²Hf–¹⁸²W, ^{235/238}U–^{207/206}Pb and ²⁴⁴Pu–¹³⁶Xe. In each case, relatively simple assumptions are made about how parent and daughter isotopes distribute themselves between the developing geochemical reservoirs, and a model age or timescale for the process of interest is calculated. The Hf–W and U–Pb chronometers, for example, yield, in principle, the time of formation of the core¹¹. In these cases, the parent elements (Hf and U) exhibit lithophile behaviour and are retained in silicate reservoirs during accretion, whereas the daughters (W and Pb) are considered to be strongly partitioned into the growing core. This partitioning enables us to calculate an apparent time of core formation assuming separation in a single step. A more realistic approach is to take account of growth of the Earth over millions of years with progressive segregation of the core as more material was added¹² (Fig. 1). In such models the growth of the Earth can be viewed as taking place at an exponentially decreasing rate, as follows (Fig. 1):

$$F_t = 1 - e^{-\lambda \Delta t} \quad (1)$$

where F_t is the cumulative fractional mass of the Earth relative to the present day at time t , λ is the time constant for accretion ($= 1/\tau$, where τ is the mean life) and $\Delta t = t_o - t$ (where t_o is the age of the Solar System).

If it is assumed that all accreting material mixed isotopically with the silicate Earth, that the accretion rate decreased exponentially with time and that the parent/daughter element ratios were fractionated in a constant fashion during core formation, a time constant for

accretion and core formation can be derived¹³. On this basis, both Hf–W and U–Pb should yield the same time constant (or its inverse, the mean life). Instead, the U–Pb timescales are more protracted than those of Hf–W (ref. 11) (Fig. 1).

One possible explanation for the disparity is that the U–Pb timescale is closer to the true Earth accretion rate, and that the short Hf–W timescale is due to incomplete equilibration of impactor metal with the silicate Earth⁵. This idea can be tested using independent estimates of the accretion rate of the Earth. The only precise estimate is provided by the time of the Moon-forming giant impact. The most widely accepted model for the origin of the Moon is that a Mars-sized object collided with the Earth at the end of its accretion, and generated the angular momentum and an Fe-depleted Moon from the resulting debris disk¹⁴. This is considered to have added the last ~10% of the Earth's mass. Therefore, the age of the Moon defines the last major growth stage for the Earth. The Hf–W age of the Moon appears to lie between 40 and 50 Myr (refs 5, 9, 11). With the Moon forming at 45 ± 5 Myr after the beginning of the Solar System, it can be shown that on average, 70–90% of the metal in the core was added under conditions of W isotopic equilibrium with the growing silicate Earth¹⁵. The small amount of disequilibrium explains why the Hf–W age appears older than is expected from the age of the Moon, but does not eliminate the discrepancy with U–Pb.

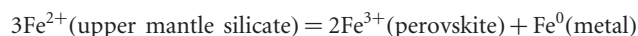
The simplest explanation for the difference in age is, of course, that Pb actually entered the Earth's core at a later stage than W. This would have required the conditions of core formation to change during the later stages of accretion. As shown below, such a change almost certainly occurred.

The partitioning of W and Pb between metals and silicates depends strongly on the oxygen fugacity of equilibrium and the sulphur content of the metal. The FeO content of the upper mantle (8 wt%), combined with the estimated Fe content of the core (80%; refs 16, 17), indicates that equilibrium core segregation would occur at $1.5\text{--}2 \log f_{O_2}$ units below the iron–wüstite (Fe–FeO) buffer¹⁸. Under these conditions, W would be a siderophile element while Pb partitioning into the metal phase would be very weak¹⁹. Currently, the upper mantle is much more oxidized, however, exhibiting oxygen fugacities about $4 \log f_{O_2}$ units higher than those needed for Fe saturation²⁰. Under these more oxidizing conditions, the only 'metallic' phase stable in the mantle is an Fe–Ni sulphide such as that found commonly in mid-ocean-ridge basalt²¹. Importantly, Pb is chalcophile, exhibiting partition coefficients of the order of 2,000 between iron sulphide liquid and iron metal²², whereas W is the opposite with partitioning values of the order of 0.01 (ref. 23). Segregation of a small amount of late sulphide liquid, the 'Hadean matte'²⁴, could therefore have been a process that fractionated U/Pb, re-setting the U–Pb age with almost no effect on W. This leads to questions of when the Earth's mantle became oxidized and whether or not such a

¹Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK. ²Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK. [†]Present address: Department of Earth and Planetary Sciences, Macquarie University, New South Wales 2109, Australia.

sulphur-rich liquid can have been extracted from it at the end of core formation.

Oxidation of the Earth's mantle is normally ascribed to dissociation of H_2O at the end of accretion, with loss of hydrogen to space²⁵. The amount of H_2O that would be required to oxidize FeO in order to reach the current Fe_2O_3 content of the mantle (~ 0.3 wt%) is extremely large, however, close to the current volume of the hydrosphere. Ascribing oxidation to this mechanism also begs the question of why Mars, a more volatile-rich planet than Earth, has a more reduced mantle than does the Earth. Recent experimental results¹⁰ demonstrate, however, that a large planet such as the Earth undergoes 'self-oxidation' at pressures and depths beyond those required to stabilize the magnesium silicate perovskite phase, $(\text{Mg,Fe})\text{SiO}_3$. The latter phase, dominant in the lower mantle (below 660 km depth) has such a strong affinity for Fe^{3+} that it forces disproportionation of Fe^{2+} as follows¹⁰:



Fe^{3+} replaces Mg^{2+} in the perovskite structure, charge-balanced by substitution of Al^{3+} for Si^{4+} . Thus, when the Earth had grown large

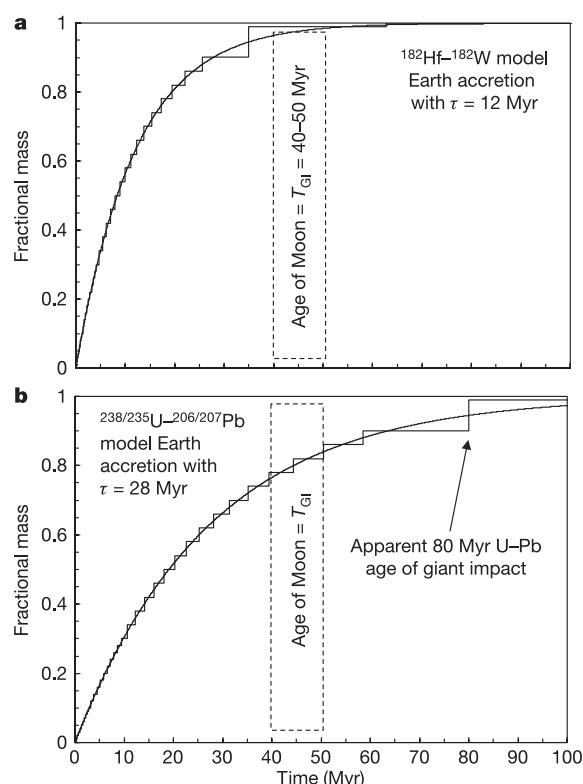


Figure 1 | Application of the growth model to the determination of the time of the Moon-forming impact. **a**, Using Hf–W chronometry; **b**, using U–Pb chronometry; the growth model is given in Methods. The tungsten isotopic composition of the silicate Earth implies a mean life of accretion (τ) of about 12 Myr and growth from 90% to 99% (due to the giant impact) about 35 Myr after the start of the Solar System. Pb data, in contrast, are consistent with a mean life of accretion of about 28 Myr, implying growth from 90% to 99% at about 80 Myr (ref. 5). A possible explanation to consider for this is that accretion was protracted, as indicated by Pb isotope data, but that isotopic equilibration between impactors and the silicate Earth was incomplete⁵. However, the Hf–W age of the Moon (T_{GI}) provides an independent test of this and is now reasonably well-defined, lying in the range 40–50 Myr (refs 5, 9, 15). Although incomplete equilibration explains the W isotopic composition of the BSE, it is inadequate as an explanation for the later apparent ages determined from Pb isotopes. A viable alternative explanation is that Pb was lost from the BSE at a relatively late stage, following formation and segregation of iron sulphide, thereby increasing the $^{238}\text{U}/^{204}\text{Pb}$ ratio of the BSE.

enough (about the size of Mars) to stabilize perovskite in the lower mantle, then, with core segregation removing metal produced by the above reaction, the $\text{Fe}^{3+}/\text{Fe}^{2+}$ ratio of the lower mantle started to increase. The vigour of accretion and core segregation would certainly have caused continuous dissolution and re-precipitation of the perovskite with release of Fe^{3+} to the overlying mantle. Given an episodically very deep ($>1,000$ km) magma ocean^{26,27}, this process would have oxidized the molten upper mantle and as recently shown²⁷, generated the observed mantle abundances of moderately siderophile elements.

Figure 2 shows the effects when ferric iron is recycled back into a melt of peridotitic composition under upper mantle conditions. Oxygen fugacity increases by about 3 log units as the Fe^{3+} content increases to the current upper mantle value. This is accompanied by destabilization of coexisting metal so that sulphur-rich liquids are the only metallic phases that can coexist with the putative magma ocean. Thus, crystallization and re-crystallization of perovskite would have raised the Fe^{3+} content of the upper mantle and terminated metal segregation from the latter. Sulphur, which had partitioned between metal and silicate, would then have begun to build in concentration in the silicate. Oxidation of later infalling metal by the ferric iron present in the magma ocean would have raised the FeO content of the mantle and left sulphur dissolved in the molten silicate.

The giant impact is believed to have added $\sim 10\%$ to the mass of the Earth and caused complete melting¹⁴, thereby releasing more Fe^{3+} through complete perovskite dissolution. At this stage, temperatures in the silicate Earth ranged from about 4,600 to 10,000 K (ref. 14), and the sulphur present before impact, together with any added to Earth from the impactor, would have been highly soluble in the liquid mantle. Note that lunar basalts are relatively rich in chalcophile elements²⁸, implying that the impactor added significant amounts of sulphide to the Earth. Sulphur solubility in silicate melt decreases strongly on cooling, leading to exsolution, in Fe-bearing systems, of an FeS -rich liquid²⁹. Thus, during cooling, an FeS -rich

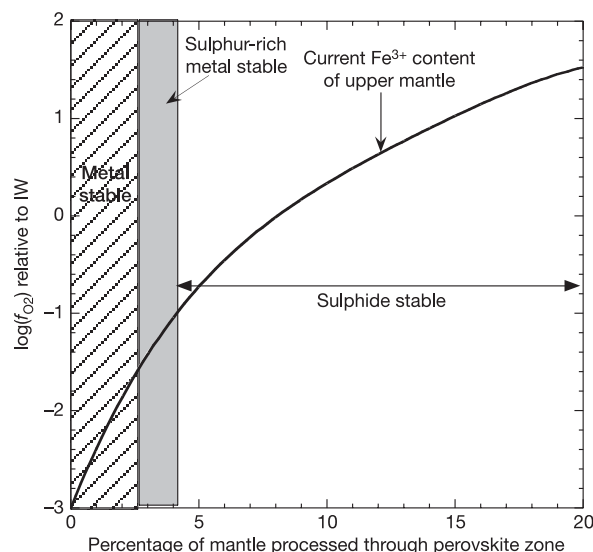


Figure 2 | Calculated effect of perovskite crystallization on the Fe^{3+} content and hence oxygen fugacity (f_{O_2}) of a magma ocean of peridotite composition. The calculation began by assuming peridotitic mantle of current composition containing 8% FeO. Perovskite crystallized from a melt of this composition contains about 3.5% 'FeO', but rather than being entirely ferrous, half of the iron is actually ferric^{10,32}. The Fe^{3+} content of the molten mantle increases as perovskite crystallizes, dissolves releasing Fe^{3+} and recrystallizes at the base of the magma ocean. This 'processing' of the mantle was converted from Fe^{3+} content to oxygen fugacity using experimental data on silicate melts³³. Oxygen fugacities are expressed relative to the iron–wüstite (IW; Fe-FeO) buffer.

liquid would have 'rained-out' of the liquid mantle, extracting chalcophile elements such as Pb and Te from the mantle. Segregation of an extremely small volume (<1%) of this sulphide matte to the core would have resulted in a young apparent U-Pb age of core formation but would have had no effect on the Hf-W age because the ^{182}Hf parent was already effectively extinct. The matte would have removed about 1,500 p.p.m. of sulphur (as FeS) from the mantle, an amount which, given the strong temperature dependence of FeS solubility in silicate melts²⁹ and the chalcophile nature of the impactor²⁸, seems reasonable.

The data summarized above suggest that Pb was lithophile during most of Earth's accretion. At a late stage, owing to oxidation of the mantle, Pb became chalcophile such that most of it was lost to the core in a sulphide liquid. The results of modelling 11 estimates of the Pb isotopic composition of the bulk silicate Earth (BSE) in such a fashion⁵, assuming the giant impact at 45 Myr, are shown in Fig. 3. The $^{238}\text{U}/^{204}\text{Pb}$ of the BSE (μ_{BSE}) is assumed to be 0.7—identical to the total Earth, until the point at which it increases owing to sulphide segregation, generating the $^{238}\text{U}/^{204}\text{Pb}$ determined from its Pb isotopic composition. The timing of this event is shown in Fig. 3 and ranges from approximately 50 to 160 Myr after the start of the Solar System. Use of the three most recent estimates defines a more restricted range from 65 to 85 Myr, 30 ± 15 Myr after a 45 ± 5 Myr

giant impact. A more conservative assumption would be to assume, since perovskite started to crystallize when the Earth was quite small, that some sulphide extraction and Pb removal pre-dated the giant impact. This assumption, when used in conjunction with the Pb isotopic composition of the BSE, has the effect of delaying the calculated time of the most recent fractionation of U from Pb. The effect is shown in Fig. 3, in which the timing of sulphide segregation is shown (open symbols) calculated following early growth of the Earth and its core with a $\mu_{\text{BSE}} (^{238}\text{U}/^{204}\text{Pb})$ of 2.0 instead of 0.7. The timing of the change to a higher μ ranges from 80 to 100 Myr (instead of 65–85 Myr) after the start of the Solar System for the three most recent estimates of the Pb isotopic composition of the BSE. Although we present its potential effects in Fig. 3, we consider it unlikely that significant amounts of 'early' sulphide were produced. This is because early accretion was accompanied by addition of substantial amounts of metal. Metal added to the molten upper mantle would have consumed ferric iron and held the oxygen fugacity below the point at which sulphide becomes stable (Fig. 2). Only after release of large amounts of Fe^{3+} from perovskite dissolution during the giant impact would the process depicted in Fig. 2 have led to a substantial increase in oxygen fugacity and sulphide precipitation. Thus the U–Pb age of the Earth (65–85 Myr after the beginning of the Solar System) most probably represents the very last stages of core segregation at the end of accretion.

The apparent time after giant impact for sulphide extraction (30 ± 15 Myr) is remarkably close to Lord Kelvin's final estimate (24 Myr)¹ for the time taken for the Earth to cool from fully molten to its current state. More recent models of convection and cooling of the Earth after a giant impact³⁰ indicate, however, that solidification of most of the mantle took only 3,000 yr. On the other hand, a shallow (50 km) slowly convecting, partially crystalline magma ocean might have existed for 10^8 yr (ref. 30). This indicates that the U–Pb age does not reflect the time taken for cooling to the point where sulphide 'rained-out' of the mantle. Rather, it must reflect the time taken to physically accumulate and extract the small volumes of sulphide involved. Thus, given the differences in physical processes involved, we conclude that the agreement of our results with Kelvin's age of the Earth is purely coincidental.

METHODS

Assumptions and model. To model the isotopic data usefully, one has to make assumptions about how accretion and core formation took place¹³. The simplest model is one in which the Earth accreted and then the core segregated. This provides a model age for the last total equilibration of the core and silicate Earth. Such ages are somewhat misleading, however, as such a single separation event almost certainly never occurred⁵. Dynamic simulations provide evidence that the Earth was in fact largely built by a stochastic series of collisions with other planetesimals and planets. The rate of accretion is expected to decrease with time, although the size of possible impacting planets would increase³¹. The model used here is identical to that of Halliday⁵ (Fig. 1). The earliest stage of growth is a runaway process that builds the Earth to 1% of its current mass over $\sim 10^5$ yr, as suggested by simulations. Further growth is dominated by collisions between these objects. The overall rate of accretionary growth of the Earth decreased with time, but the growth events would have become more widely interspersed and larger. Therefore, the model simulates further growth by successive additions of 1% Earth mass objects from 1% to 10% of the current mass, then by 2% objects to 30% and then by 4% objects to 90%. The Moon-forming giant impact is modelled as taking place when the Earth was $\sim 90\%$ of its current mass, and involved an impactor planet Theia that was $\sim 10\%$ of the (then) mass of the Earth (Fig. 1). Therefore, the giant impact contributes a further 9% of the current Earth mass. A late veneer contributes an additional 0.9%. There is a final 0.1% added after this. The same accretion timescale can therefore be expressed in two ways—as a model time of the giant impact (that is, the time at which the Earth grew from 90% to 99% of the current mass), and as a mean life of accretion for exponentially decreasing growth.

Calculations. All calculations in this Letter assume continuous core formation and total equilibration between accreted material and the BSE. The partition coefficients for U and Pb between the core and the silicate Earth were varied, as detailed in the text. All $^{238}\text{U}/^{204}\text{Pb}$ (or μ) values are present day equivalent values. The μ for the total Earth ($\mu_{\text{TOTE}} = 0.7$). The present day ($^{206}\text{Pb}/^{204}\text{Pb}$)_{BSE} can be

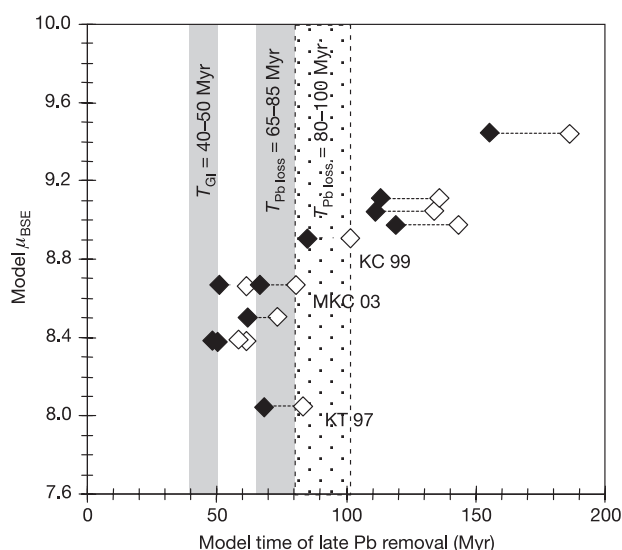


Figure 3 | Results of modelling 11 estimates of the Pb isotopic composition of the BSE with late stage loss of Pb⁵. It is assumed that Pb remained lithophile until a point in time when the majority was instantaneously partitioned into the core to achieve the $^{238}\text{U}/^{204}\text{Pb}$ or μ of the BSE plotted on the y axis. This is the μ defined by the single-stage Pb isotopic age of the Earth given by the present Pb isotopic composition. The exact μ value is not very sensitive to the accretion and partitioning history because it is provided by $^{206}\text{Pb}/^{204}\text{Pb}$, reflecting decay of long-lived ^{238}U . The accretion history and exact time of Pb segregation is more sensitive to $^{207}\text{Pb}/^{204}\text{Pb}$ dominated by early and rapid decay of relatively short-lived ^{235}U . It is assumed the giant impact occurred at 45 Myr after the beginning of the Solar System, defining the timing of growth of the Earth from 90% to 99% of its present mass. It can be seen that the model age for the timing of the increase in μ of the BSE postdates the 45 Myr age of the giant impact. The overall range in ages is broad but much narrower (65–85 Myr) for the three most recent estimates of the Pb isotopic composition of the BSE (KT97, KC99 and MKC03 as referenced in ref. 5). This approximates the time when, we suggest, sulphide was extracted from the silicate Earth. If lead were in fact partly siderophile during early Earth accretion, the timing of sulphide formation would be even later. The open symbols show the results for the same calculation assuming that before sulphide segregation lead was partly siderophile, leading to a μ for the BSE of 2.0 instead of 0.7. The timescales are increased somewhat, the three most recent estimates of the Pb isotopic composition of the BSE yielding values of 80–100 Myr.

used to place limits on the current μ_{BSE} because the age of the Earth introduces a trivial uncertainty and early changes in μ are insignificant for present day $(^{206}\text{Pb}/^{204}\text{Pb})_{\text{BSE}}$. The single stage Pb–Pb model age was used to determine this. Note, however, that present day μ_{BSE} is somewhat uncertain because of the fractionation of U from Pb during geochemical processes, and that a variety of estimates are available in the literature. A range of model ages of sulphide segregation are obtained, as shown in Fig. 3. In every model, the Bulk Solar System Initial (BSSI) values are $(^{206}\text{Pb}/^{204}\text{Pb})_{\text{BSSI}} = 9.307$ and $(^{207}\text{Pb}/^{204}\text{Pb})_{\text{BSSI}} = 10.294$, with $^{238}\text{U}/^{235}\text{U} = 137.88$, and U decay constants $\lambda^{235} = 9.85 \times 10^{-10} \text{ yr}^{-1}$, $\lambda^{238} = 1.55 \times 10^{-10} \text{ yr}^{-1}$.

Received 6 April; accepted 9 August 2005.

- Kelvin, The age of the Earth. *Nature* **51**, 438–440 (1895).
- Carlsaw, H. S. & Jaeger, J. C. *Conduction of Heat in Solids* (Oxford Univ. Press, London, 1959).
- Cameron, A. G. W. & Benz, W. Origin of the Moon and the single impact hypothesis IV. *Icarus* **92**, 204–216 (1991).
- Canup, R. M. & Asphaug, E. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412**, 708–712 (2001).
- Halliday, A. N. Mixing, volatile loss and compositional change during impact-driven accretion of the Earth. *Nature* **427**, 505–509 (2004).
- Kleine, T., Munker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf–W chronometry. *Nature* **418**, 952–955 (2002).
- Yin, Q. Z. *et al.* A short timescale for terrestrial planet formation from Hf–W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
- Schoenberg, R., Kamber, B. S., Collerson, K. D. & Eugster, O. New W-isotope evidence for rapid terrestrial accretion and very early core formation. *Geochim. Cosmochim. Acta* **66**, 3151–3160 (2002).
- Kleine, T., Palme, H. & Mezger, K. The Hf–W age of the lunar magma ocean. *Proc. Lunar Planet. Sci. Conf. XXXVI*, 1940 (2005).
- Frost, D. J. *et al.* Experimental evidence for the existence of iron-rich metal in the Earth's lower mantle. *Nature* **428**, 409–412 (2004).
- Halliday, A. N. The origin and earliest history of the Earth. In *Meteorites, Comets and Planets* (ed. Davis, A. M.) 509–557 (Elsevier-Pergamon, Oxford, 2003).
- Stevenson, D. J. Models of the Earth's core. *Science* **214**, 611–619 (1981).
- Harper, C. L. Jr & Jacobsen, S. B. Evidence for ^{182}Hf in the early Solar System and constraints on the timescale of terrestrial accretion and core formation. *Geochim. Cosmochim. Acta* **60**, 1131–1153 (1996).
- Canup, R. M. Simulations of a late lunar-forming impact. *Icarus* **168**, 433–456 (2004).
- Halliday, A. N. & Kleine, T. Meteorites and the timing, mechanisms and conditions of terrestrial planet accretion and early differentiation. In *Meteorites and the Early Solar System II* (eds Lauretta, D., Leshin, L. & MacSween, H.) (Univ. Arizona Press, Tucson, in the press).
- McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120**, 223–253 (1995).
- Allegre, C. J., Poirier, J.-P., Humler, E. & Hofmann, A. W. The chemical composition of the Earth. *Earth Planet. Sci. Lett.* **134**, 515–526 (1995).
- Righter, K. & Drake, M. J. Metal-silicate equilibrium in a homogeneously accreting earth: New results for Re. *Earth Planet. Sci. Lett.* **146**, 541–553 (1997).
- Ohtani, E., Yurimoto, H. & Seto, S. Element partitioning between metallic liquid, silicate liquid, and lower-mantle minerals: implications for core formation of the Earth. *Phys. Earth Planet. Inter.* **100**, 97–114 (1997).
- Wood, B. J., Bryndzia, L. T. & Johnson, K. E. Mantle oxidation state and its relationship to tectonic environment and fluid speciation. *Science* **248**, 337–345 (1990).
- Francis, R. D. Sulfide globules in mid-ocean ridge basalts (MORB), and the effect of oxygen abundance in Fe–S–O liquids on the ability of those liquids to partition metals from MORB and komatiite magmas. *Chem. Geol.* **85**, 199–213 (1990).
- Jones, J. H., Hart, S. R. & Benjamin, T. M. Experimental partitioning near the Fe–FeS eutectic, with an emphasis on elements important to iron meteorite chronologies (Pb, Ag, Pd, and Ti). *Geochim. Cosmochim. Acta* **57**, 453–460 (1993).
- Chabot, N. L. & Jones, J. H. The parameterization of solid metal-liquid metal partitioning of siderophile elements. *Meteorit. Planet. Sci.* **38**, 1425–1436 (2003).
- O'Neill, H. S. The origin of the Moon and the early history of the Earth—A chemical model. Part 2: The Earth. *Geochim. Cosmochim. Acta* **55**, 1159–1172 (1991).
- Kasting, J. F., Egglar, D. H. & Raeburn, S. P. Mantle redox evolution and the oxidation-state of the Archean atmosphere. *J. Geol.* **101**, 245–257 (1993).
- Chabot, N. L., Draper, D. S. & Agee, C. B. Conditions of core formation in the Earth: Constraints from nickel and cobalt partitioning. *Geochim. Cosmochim. Acta* **69**, 2141–2151 (2005).
- Wade, J. & Wood, B. J. Core formation and the oxidation state of the Earth. *Earth Planet. Sci. Lett.* **236**, 78–95 (2005).
- Yi, W. *et al.* Cadmium, indium, tin, tellurium, and sulfur in oceanic basalts: Implications for chalcophile element fractionation in the Earth. *J. Geophys. Res. Solid Earth* **105**, 18927–18948 (2000).
- Holzheid, A. & Grove, T. L. Sulfur saturation limits in silicate melts and their implications for core formation scenarios for terrestrial planets. *Am. Mineral.* **87**, 227–237 (2002).
- Abe, Y. Thermal and chemical evolution of the terrestrial magma ocean. *Phys. Earth Planet. Inter.* **100**, 27–39 (1997).
- Wetherill, G. W. Accumulation of the terrestrial planets and implications concerning lunar origin. In *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. J.) (Lunar and Planetary Institute, Houston, 1986).
- Tronnes, R. G. & Frost, D. J. Peridotite melting and mineral-melt partitioning of major and minor elements at 22–24.5 GPa. *Earth Planet. Sci. Lett.* **197**, 117–131 (2002).
- Kilinc, A., Carmichael, I. S. E., Rivers, M. L. & Sack, R. O. The ferric-ferrous ratio of natural silicate liquids equilibrated in air. *Contrib. Mineral. Petrol.* **83**, 136–140 (1983).

Acknowledgements Discussions with J. Wade and M. Walter helped to clarify our arguments. B.J.W. acknowledges the support of a Max-Planck Research Award.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.J.W. (bwood@els.mq.edu.au).

Mammal-like muscles power swimming in a cold-water shark

Diego Bernal^{1,2}, Jeanine M. Donley³, Robert E. Shadwick^{2†} & Douglas A. Syme⁴

Effects of temperature on muscle contraction and powering movement are profound, outwardly obvious, and of great consequence to survival^{1,2}. To cope with the effects of environmental temperature fluctuations, endothermic birds and mammals maintain a relatively warm and constant body temperature, whereas most fishes and other vertebrates are ectothermic and conform to their thermal niche, compromising performance at colder temperatures^{2,3}. However, within the fishes the tunas and lamnid sharks deviate from the ectothermic strategy, maintaining elevated core body temperatures^{4,5} that presumably confer physiological advantages for their roles as fast and continuously swimming pelagic predators. Here we show that the salmon shark, a lamnid inhabiting cold, north Pacific waters, has become so specialized for endothermy that its red, aerobic, locomotor muscles, which power continuous swimming, seem mammal-like, functioning only within a markedly elevated temperature range (20–30 °C). These muscles are ineffectual if exposed to the cool water temperatures, and when warmed even 10 °C above ambient they still produce only 25–50% of the power produced at 26 °C. In contrast, the white muscles, powering burst swimming, do not show such a marked thermal dependence and work well across a wide range of temperatures.

An anatomical feature that sets the lamnid sharks apart from ectothermic fishes but reveals evolutionary convergence with tunas is the location of the red, aerobic, locomotor muscle (RM), being deep in the body next to the vertebral column and concentrated in the mid-body region, in contrast with the lateral and subcutaneous position of RM in other fishes⁶. Coupled with vascular heat exchangers, tunas and lamnid sharks have evolved the ability to retain metabolic heat in RM and attain temperatures that are significantly above that of the surrounding water^{5,6}. Furthermore, the white locomotor muscle (WM) is also significantly warmed by conductive heat transfer arising from its proximity to the RM, with progressive cooling towards the skin⁷. Thus, two of the most striking differences between endothermic and ectothermic fishes are the position of the RM and the capacity to warm both the RM and a portion of the WM.

What are the functional consequences of these thermal and anatomical modifications? Previous work on the swimming physiology and metabolic biochemistry of muscles in warm-bodied lamnid sharks^{7,8} and tunas^{9,10} suggests that these fishes have the potential to sustain a higher aerobic swimming metabolism, and have an increased potential for burst swimming relative to that of other fishes. However, there is no direct experimental evidence demonstrating how a change in the operating temperature of these tissues might affect the contractile properties of shark muscle, if the maintenance of high temperatures in RM is necessary for

satisfactory function, or if warm-bodied sharks are capable of enhanced swimming performance¹¹. To address these questions we examined thermal effects on the contractile properties of RM and WM in salmon sharks (*Lamna ditropis*), a species inhabiting the cold waters of the north Pacific Ocean¹² but noted for maintaining elevated body temperatures^{6,13,14} and prized as a sport-fish for its impressive swimming power and speed¹⁵.

Within minutes of landing the sharks we mapped the thermal gradient in the muscle mass by measuring temperatures along the body. Measurements taken just anterior to the first dorsal fin, where the RM is most abundant, showed that all sharks had a core temperature about 18–20 °C above the surrounding sea surface temperature, which ranged from 6 °C to 8 °C; this gradient is even larger than those reported previously^{6,13,14}, probably due to the somewhat cooler surface temperature at the time of this study. Regardless of variations in surface temperatures, tagging studies have shown that salmon sharks routinely dive to 200 m and encounter water temperatures of 6 °C during most of the year¹². The maximum temperature of the deeply positioned RM was 26 °C, which was 16 °C warmer than the superficial WM at this position and about 20 °C warmer than the ambient water. A smaller thermal gradient was present at the level of the second dorsal fin, a narrower region of the body, with the deep RM temperature being 10–13 °C above ambient. In Fig. 1, the thermal data from all specimens are shown superimposed on a three-dimensional reconstruction of the distribution of locomotor muscle in the salmon shark, providing a perspective of the operating temperatures *in situ* for the RM and WM along both the longitudinal and transverse axes of the body. Temperatures selected from this reconstruction were used to determine thermal effects on muscle contractile properties.

Using a portable stimulator/transducer to make measurements of muscle twitches *in situ*, we next determined that the duration of twitches from superficial WM was invariant along the length of the body (see Supplementary Fig. 1) but decreased significantly (that is, the twitches got faster) with depth along the temperature gradient of WM from the cool skin to the warm body core. Contractile tests were then conducted on small bundles of living muscle fibres isolated from a midbody position (that is, at the level just anterior to the first dorsal fin), where the RM and WM temperatures were the most elevated. Durations of isometric twitches measured from WM ranged from 143 ms at 10 °C to 51 ms at 26 °C, yielding an average thermal rate coefficient (Q_{10}) of 2.0 (Fig. 2), a typical temperature response for muscle from an ectothermic vertebrate¹. In contrast to WM, the twitch times of RM had both much higher Q_{10} values (up to 3.7) and exceedingly long twitch durations, with a time to reach peak force exceeding 3 s at 10 °C, and even longer relaxation times (see Fig. 2). These observations indicate that the RM is extremely sensitive

¹Department of Biology, University of Massachusetts, Dartmouth, North Dartmouth, Massachusetts 02747, USA. ²Marine Biology Research Division, Scripps Institution of Oceanography, La Jolla, California 92093-0202, USA. ³Department of Biological Sciences, Miracosta College, Oceanside, California 92056, USA. ⁴Department of Biological Sciences, University of Calgary, Alberta, T2N 1N4 Canada. [†]Present address: Department of Zoology, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada.

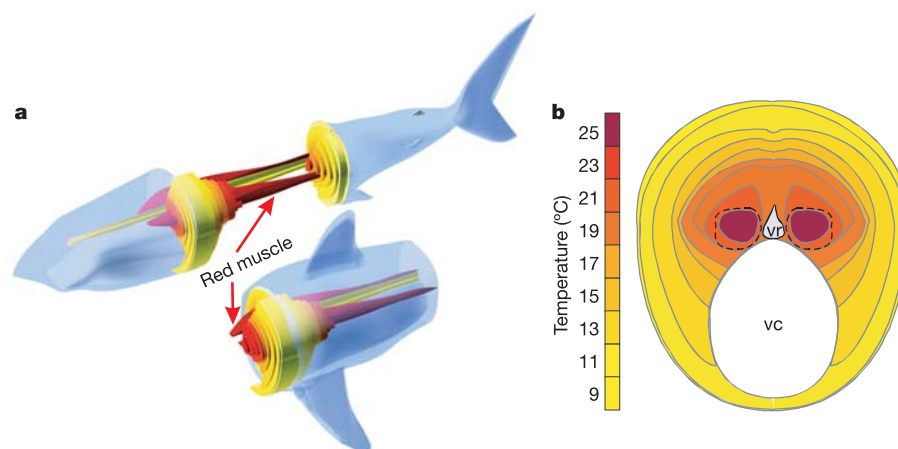


Figure 1 | Thermal gradients superimposed on a three-dimensional map of RM and WM distribution in salmon sharks (*Lamna ditropis*). **a**, Isotherms shown are the mean value for two specimens (216 and 222 cm FL, about 140 and 160 kg body mass) at two positions along the body (40% and 63% of FL). **b**, The body section just anterior to the first dorsal fin (lower section in

a) that contains the highest cross-sectional area of RM. Dashed circles in transverse images indicate red muscle position. The external surface is at the same temperature as the water, 8.5 °C. Abbreviations: vc, visceral mass; vr, vertebrae.

to temperature, particularly at cooler temperatures, and would probably be incapable of contracting in a way that would result in effective swimming if allowed to cool even to 15 °C (a temperature still 9 °C above ambient).

To understand how temperature might limit the ability of these muscles to power swimming, we characterized their power-producing potential by using work-loop analysis¹⁶ at temperatures encompassing those measured in these sharks and over a range of cycle (that is, tail-beat) frequencies that they might use during swimming. The temperature-dependent power spectra of WM indicate a retained ability to produce high power at tail-beat frequencies of 4–6 Hz between temperatures of 10 °C and 26 °C, with power output being maintained up to a frequency of 8 Hz at 26 °C (Fig. 3a). Thus, WM of salmon sharks shows a thermal sensitivity typical of that observed in most other temperate and cold-water fishes^{2,17}, allowing WM to function readily at temperatures between 10 and 26 °C, well within the 18–20 °C thermal gradient noted between the cold surface and the warm interior of the shark's body. In marked contrast, the fastest tail-beat frequency at

which the RM produced power was restricted to a relatively slow 2 Hz even at a very warm temperature of 31 °C, with power production being very low and further limited to a tail-beat frequency of only 0.5 Hz at 15 °C (Fig. 3b). The RM of salmon sharks seems to be designed for high power production only at very high body temperatures (26 °C or above) and could not produce useful work and power if allowed to cool below about 20 °C (Fig. 3b), a mere 6 °C below its operating temperature *in vivo* but still considerably more than 10 °C warmer than the water. Thus, the superficial WM can function well at cold temperatures, but like mammalian skeletal muscle¹ the internalized RM has become an obligate endothermic tissue and functions only at high body temperatures. Further, the elevated temperature of WM close to the warm RM near the body core would result in an approximately threefold increase in its ability to produce power, probably further enhancing the burst swimming capabilities of these fish.

Our results show that WM from salmon sharks can effectively power burst swimming at relatively high tail-beat frequencies across a wide range of temperatures; it is therefore well-suited to work in the

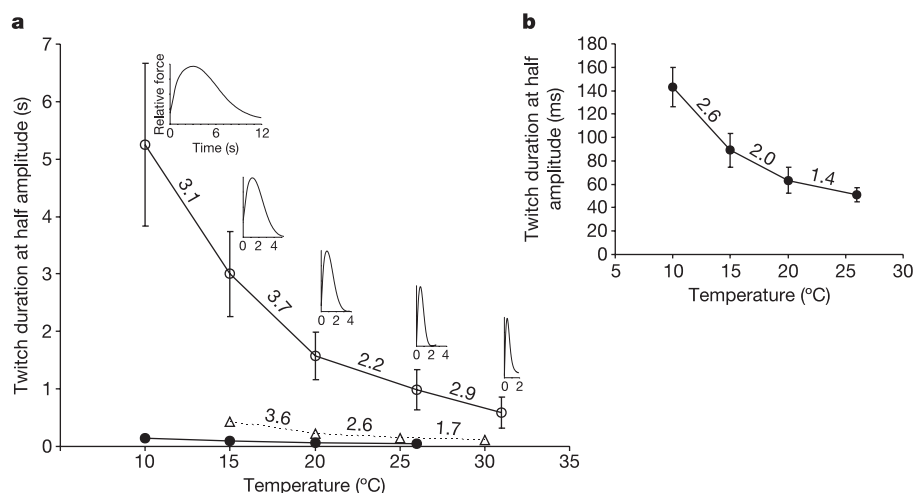


Figure 2 | Muscle twitch duration at different temperatures. **a**, Twitch in RM (open circles) and WM (filled circles) from *Lamna ditropis* and in RM from *Thunnus albacares* (open triangles). The thermal rate coefficients (Q_{10}) for twitch duration are shown at each temperature interval. Insets show single-twitch profiles at each temperature for RM from *L. ditropis*.

b, Twitch durations for WM from *L. ditropis*, with expanded axes for clarity. Values shown are means $\pm$ s.d. Each point of twitch data for salmon sharks is the period for which the force was greater than 50% of maximal, and for RM from yellowfin tuna (*T. albacares*) it is the time from stimulus to half relaxation¹⁸.

cold but probably realizes a considerable benefit in being warmed through its close proximity to the RM. In contrast, the RM used to power sustained swimming is specialized for use at warm temperatures only, perhaps minimally 15 °C warmer than the water that the fish inhabits, and the shark's highly endothermic nature is essential to the functioning of these muscles. In this sense the RM of salmon sharks is analogous to that of mammals in functioning only at relatively warm temperatures and having Q_{10} values for twitch speed that are higher than those typically observed in ectothermic muscles and are more similar to those for endothermic muscle¹. This tissue is unlike muscle found in any other fish so far, including that of the endothermic tunas. Both salmon sharks and tunas seem to be similar in having very thermally sensitive RM, particularly in the deeper, warmer regions of the fish¹⁸, but differ in that the twitches of tuna RM are about sixfold faster¹⁸ than those noted in salmon sharks at similar temperatures (Fig. 2). Further, unlike salmon sharks, tuna RM can continue to power relatively fast swimming (for example with a tail-beat frequency of 2–4 Hz) even at a cool temperature of 15 °C (ref. 18). Whereas the large size of the salmon sharks may contribute in part to the slow contractions of its RM, evidence from dogfish sharks¹⁹ and tuna RM¹⁸ indicate that scaling with body size might not be an important influence, and even after correcting for potential effects of body size using scaling coefficients from studies on other species²⁰, the twitch speeds of RM in salmon sharks remain about fourfold slower than those from RM of tuna.

Unlike mammals and birds, which do not depend solely on locomotor activity to warm their muscles and can use other sources of metabolic heat during times of inactivity, sharks have no known non-locomotor-associated thermogenic tissues and therefore salmon

sharks must rely on the constant contractile activity of their RM to elevate this tissue's temperature. Thus, these sharks live in a physiological extreme in that if the aerobic RM were to become inactive (that is, if the fish stopped swimming) and ceased to produce heat, allowing the temperature to fall below even 20 °C, which is still about 15 °C warmer than the surrounding water, the RM would fail to function and possibly could not recover. Obligate, continuous locomotion therefore seems to be essential in allowing these sharks to maintain warm RM temperatures that in turn permit them to exploit, all year round, the north Pacific's energy-rich cold waters as apex pelagic predators.

METHODS

Fish collection. Three salmon sharks (*Lamna ditropis*) were captured by hook and line in May and June off Prince William Sound in the Gulf of Alaska, USA. Specimens ranged from 2.16 to 2.22 m fork length (FL), with a body mass ranging from about 130 kg to 160 kg. All capture and experimental procedures followed the guidelines of the University of California, San Diego IACUC.

In situ temperature and twitch measurements. A thermocouple was used to measure tissue temperatures at several positions along the body in freshly caught specimens; recordings began within 60 s of landing. Two small incisions were made at 45% of FL, the first just anterior to the first dorsal fin and the second along the side of the body between the epaxial and hypaxial musculature, and a thermocouple probe was advanced in an orthogonal direction towards the vertebrae, eventually reaching the spine. Temperature data were recorded every 10 mm along these trajectories. Similar measurements were obtained near the second dorsal fin; thermal data for both positions along the body were mapped on a three-dimensional model of salmon shark locomotor muscles and isotherms were interpolated at intervals of 2 °C.

Muscle preparation. Segments of RM and WM (about 1 cm³) were removed from directly under the first dorsal fin (about 45% of FL) shortly after capture. A small bundle of fibres about 1 mm in diameter and spanning several myotomes in length was then isolated under a dissecting microscope on a cooled stage (about 4 °C) with frequent rinses of chilled shark saline (292 mM NaCl, 3.2 mM KCl, 5 mM CaCl₂, 0.6 mM MgSO₄, 1.6 mM Na₂SO₄, 300 mM urea, 150 mM trimethylamine N-oxide, 10 mM glucose, 6 mM NaHCO₃; total osmolality 1,080 mosM, pH 7.7 at 20 °C). These preparations were immersed in oxygenated saline on ice during transport back to the laboratory (about 180 min). Muscle bundles were then further dissected to remove damaged fibres and reduced to a single myotome. Preparations were transferred to a temperature-controlled experimental chamber filled with shark saline; the saline was circulated through a temperature-controlled reservoir and bubbled with pure oxygen. The myoseptum on one end of the bundle was tied to a stainless-steel pin connected to a servomotor arm (model 350 for RM and model 305B for WM; Cambridge Technology) and the myoseptum on the other end was tied to a pin connected to a force transducer (a model BG-50G (Kulite Semiconductor Products Inc.) for RM and a model FT03C (Grass Telefactor) for WM). Muscle length was adjusted to remove visible slack. Platinum-tipped electrodes or plates were positioned alongside the muscle and the stimulus voltage was adjusted to 150% of that required to elicit a maximal isometric twitch (1-ms stimulus pulse). Muscle length was then adjusted systematically until maximal twitch force was attained. All data were recorded at 1 kHz on National Instruments PCI MIO 16-E4 analogue-to-digital cards using custom software written in LabView (National Instruments). Isometric twitches were recorded over the range of temperatures as noted below. Twitch duration was measured as the period for which force was greater than 50% of the maximum recorded.

Recording work and power. Mechanical work and power were measured with the work-loop method¹⁶, in which muscle preparations were subjected to sinusoidal strain centred on the length at which force was maximal. The frequency of the strain oscillations was varied as noted below, and the amplitude of the muscle strain oscillation was fixed at $\pm 5\%$ of the reference length, similar to that recorded *in vivo* from RM of swimming tuna²¹ and mako sharks²². Preparations were stimulated by means of platinum electrodes during the strain cycle, and the timing of the stimulus was adjusted such that force was produced primarily during shortening and the muscle was relaxed during lengthening. Net work done by the muscle was calculated as the integral of force with respect to change in muscle length over a complete cycle, and power was calculated as the product of work done per cycle and cycle frequency. At each combination of cycle frequency and temperature, the stimulus parameters were adjusted to maximize the net work done during the length cycles. Work was measured over a range of cycle frequencies (equivalent to tail-beat frequencies) encompassing 0.5–3 Hz for RM and 2–7 Hz for WM. Isometric twitches were recorded

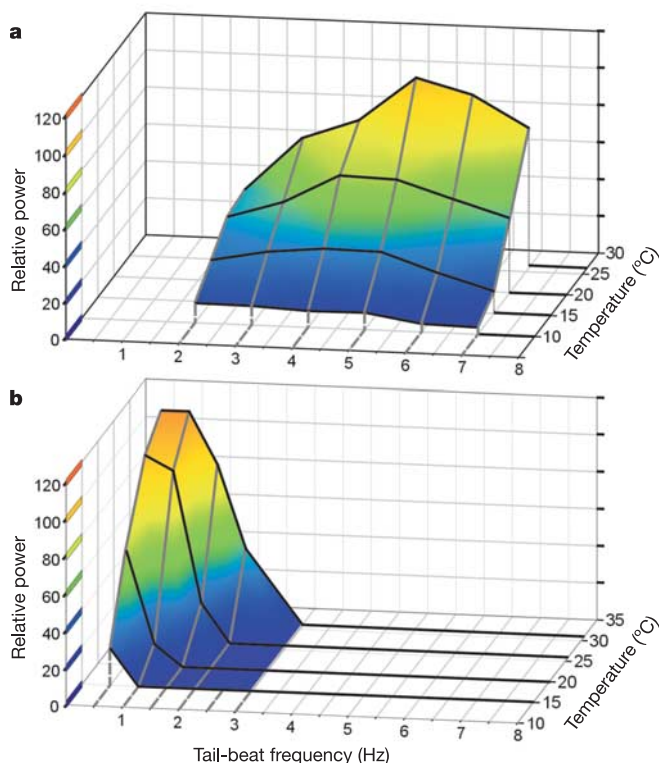


Figure 3 | Relative power output in myotomal muscle of *Lamna ditropis* at different cycle (that is, tail-beat) frequencies and temperatures, measured by work-loop analysis. a, WM; b, RM. Amplitude of muscle strain oscillation was fixed at $\pm 5\%$ of sample reference length. Stimulation parameters were optimized at each combination of temperature and cycle frequency to maximize net power output. Note the limited frequency range over which red muscle can produce power, and its reliance on high temperatures.

systematically throughout the experiments to monitor the stability of the preparations, and any muscles that exhibited a marked decline in force were discarded (this occurred in only one preparation).

Temperature effects. Isometric twitches, work and power were recorded in muscle preparations exposed to temperatures of 10, 15, 20 and 26 °C for WM and 15, 20, 26 and 31 °C for RM; twitches were also recorded at 10 °C in two preparations of RM. These temperature ranges encompassed values recorded in freshly caught specimens, with additional measurements at the lower and higher ends to determine the full extent of the thermal effects on muscle contractile properties.

Received 15 April; accepted 11 July 2005.

- Bennett, A. F. Temperature and muscle. *J. Exp. Biol.* **115**, 333–344 (1985).
- Johnston, I. A. & Temple, G. K. Thermal plasticity of skeletal muscle phenotype in ectothermic vertebrates and its significance for locomotory behaviour. *J. Exp. Biol.* **205**, 2305–2322 (2002).
- Bennett, A. F. Thermal dependence of locomotor capacity. *Am. J. Physiol.* **259**, R253–R258 (1990).
- Carey, F. G. & Teal, J. M. Mako and porbeagle: warm bodied sharks. *Comp. Biochem. Physiol.* **28**, 199–204 (1969).
- Carey, F. G. & Teal, J. M. Regulation of body temperature by the bluefin tuna. *Comp. Biochem. Physiol.* **28**, 205–213 (1969).
- Bernal, D., Dickson, K. A., Shadwick, R. E. & Graham, J. B. Analysis of the evolutionary convergence for high performance swimming in lamnid sharks and tunas. *Comp. Biochem. Physiol.* **129**, 695–726 (2001).
- Bernal, D. *et al.* Comparative studies of high performance swimming in sharks. II. Metabolic biochemistry of locomotor and myocardial muscle in endothermic and ectothermic sharks. *J. Exp. Biol.* **206**, 2845–2857 (2003).
- Graham, J. B., Dewar, H., Lai, N. C., Lowell, W. R. & Arce, S. M. Aspects of shark swimming performance determined using a large water tunnel. *J. Exp. Biol.* **151**, 175–192 (1990).
- Dewar, H. & Graham, J. B. Studies of tropical tuna swimming performance in a large water tunnel. I. Energetics. *J. Exp. Biol.* **192**, 13–31 (1994).
- Dickson, K. A. Unique adaptations of the metabolic biochemistry of tunas and billfishes for life in the pelagic environment. *Environ. Biol. Fish.* **42**, 65–97 (1995).
- Katz, S. L. Design of heterothermic muscle in fish. *J. Exp. Biol.* **205**, 2251–2266 (2002).
- Hulbert, L. B. & Rice, S. D. *Salmon Shark, Lamna ditropis, Movements, Diet, and Abundance in the Eastern North Pacific Ocean and Prince William Sound, Alaska. Exxon Valdez Oil Spill Restoration Project Final Report* (Restoration Project 02396, NOAA Fisheries Auke Bay Laboratory, Juneau, Alaska, 2002).
- Rhodes, D. & Smith, R. Body temperature of the salmon shark, *Lamna ditropis*. *J. Mar. Biol. Assoc. UK* **63**, 243–244 (1983).
- Anderson, S. A. & Goldman, K. J. Temperature measurements from salmon sharks, *Lamna ditropis*, in Alaskan waters. *Copeia* **2001**, 794–796 (2001).
- Paust, B. & Smith, R. *Salmon Shark Manual. The Development of a Commercial Salmon Shark, Lamna ditropis, Fishery in the North Pacific* (Report No. 86–01, University of Alaska, Sea Grant College Program, Fairbanks, Alaska, 1989).
- Syme, D. A. & Shadwick, R. E. Effects of longitudinal body position and swimming speed on mechanical power of deep red muscle from skipjack tuna (*Katsuwonus pelamis*). *J. Exp. Biol.* **205**, 189–200 (2002).
- Johnson, T. P. & Johnston, I. A. Temperature adaptation and the contractile properties of live muscle fibres from teleost fish. *J. Comp. Physiol. Biochem.* **161**, 27–36 (1991).
- Altringham, J. D. & Block, B. A. Why do tuna maintain elevated slow muscle temperatures? Power output of muscle isolated from endothermic and ectothermic fish. *J. Exp. Biol.* **200**, 2617–2627 (1997).
- Curtin, N. A. & Woledge, R. C. Power output and force-velocity relationship of live fibres from white myotomal muscle of the dogfish, *Scyliorhinus canicula*. *J. Exp. Biol.* **140**, 187–197 (1988).
- James, R. S., Cole, N. J., Davies, M. L. F. & Johnston, I. A. Scaling of intrinsic contractile properties and myofibrillar protein composition of fast muscle in the fish *Myoxocephalus scorpius* L. *J. Exp. Biol.* **201**, 901–912 (1998).
- Katz, S. L., Syme, D. A. & Shadwick, R. E. High-speed swimming: Enhanced power in yellowfin tuna. *Nature* **410**, 770–771 (2000).
- Donley, J. M., Sepulveda, C. A., Konstantinidis, P., Gemballa, S. & Shadwick, R. E. Convergent evolution in mechanical design of lamnid sharks and tunas. *Nature* **429**, 61–65 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff at the University of Alaska Seward Marine Center for the use of laboratory facilities, the captain and crew of the F/V *Legend* for their assistance in the fishing operations, and J. Valdez for logistical support. This work was supported by funding from the NSF and the University of California San Diego Academic Senate (R.E.S. and D.B.) and the NSERC (D.A.S.).

Author Contributions All authors contributed equally to project planning, experimental work and writing of this Letter.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.E.S. (shadwick@zoology.ubc.ca) or D.B. (dbernal@umassd.edu).

Reinforcement drives rapid allopatric speciation

Conrad J. Hoskin¹, Megan Higgie¹, Keith R. McDonald² & Craig Moritz³

Allopatric speciation results from geographic isolation between populations. In the absence of gene flow, reproductive isolation arises gradually and incidentally as a result of mutation, genetic drift and the indirect effects of natural selection driving local adaptation^{1–3}. In contrast, speciation by reinforcement is driven directly by natural selection against maladaptive hybridization^{1,4}. This gives individuals that choose the traits of their own lineage greater fitness, potentially leading to rapid speciation between the lineages^{1,4}. Reinforcing natural selection on a population of one of the lineages in a mosaic contact zone could also result in divergence of the population from the allopatric range of its own lineage outside the zone^{4–6}. Here we test this with molecular data, experimental crosses, field measurements and mate choice experiments in a mosaic contact zone between two lineages of a rainforest frog. We show that reinforcing natural selection has resulted in significant premating isolation of a population in the contact zone not only from the other lineage but also, incidentally, from the closely related main range of its own lineage. Thus we show the potential for reinforcement to drive rapid allopatric speciation.

Although the role of reinforcing natural selection in driving allopatric speciation within a lineage has received little attention^{4,6}, its role in the final stages of speciation between divergent sister lineages has had a long and contentious history—some consider that there are compelling examples⁷, whereas others argue that the empirical evidence is generally inconclusive^{8,9}. Uncertainty remains because strict criteria must be satisfied to demonstrate the process of reinforcement unambiguously⁴: first, heterospecific matings occur in the field; second, there is selection against hybridization; third, displacement of a trait is perceived by the other sex; fourth, variation in the trait is heritable and responds to selection; and last, displacement has not occurred for other reasons, such as ecological divergence. Demonstrating that this process can complete speciation requires showing that it has resulted in significant reproductive isolation between the lineages. Further, evidence for reinforcement is most compelling within a well-corroborated historical biogeographic framework¹⁰.

The green-eyed tree-frog *Litoria genimaculata* is a stream-breeding hylid frog common in rainforest throughout the 'Wet Tropics' region of northeast Queensland, Australia. Like many other Wet Tropics

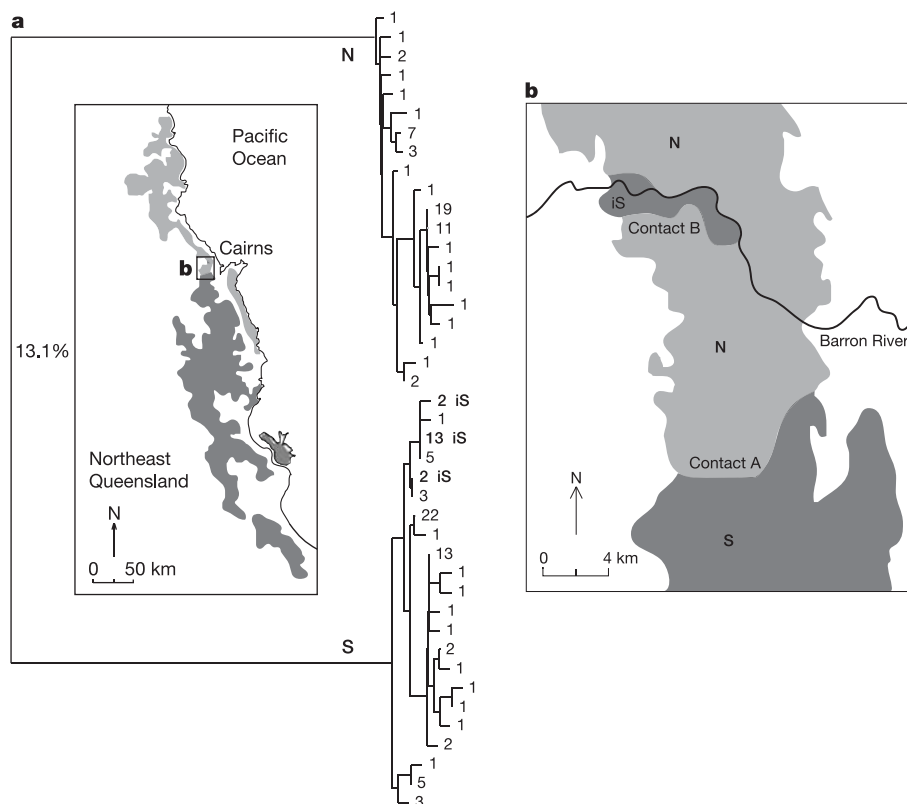


Figure 1 | Distribution of the N (pale shading) and S (dark shading) lineages of *L. genimaculata*. **a**, The distribution of N and S and a tree showing genetic divergence (percentage) and the number of individuals with each haplotype. **b**, An enlargement of the region within the rectangle in **a**, showing the mosaic contact between the N and S lineages. The iS population shares haplotypes (with 0.1% net divergence overall) with nearby populations from the main range of the S lineage (**a**) but is currently isolated from them (population subdivision statistic $F_{st} = 0.20$, $P < 0.001$). By contrast, the N populations are continuous and genetically connected (north versus south of Barron River; $F_{st} = -0.01$, $P = 0.60$) through the mosaic contact.

¹School of Integrative Biology, University of Queensland, St Lucia, Queensland 4072, Australia. ²Queensland Parks and Wildlife Service, PO Box 975, Atherton, Queensland 4883, Australia. ³Museum of Vertebrate Zoology, University of California, Berkeley, California 94720, USA.

species, *L. genimaculata* consists of two highly divergent lineages, northern (N) and southern (S) (Fig. 1a), reflecting long-term isolation to northern and southern rainforest refugia during the cooler, drier periods of the Pliocene and Pleistocene epochs¹¹. The rainforests of the northern and southern refugia reconnected about 6,500 years ago, bringing the *L. genimaculata* lineages into secondary contact in a suture zone¹². The secondary contact zone in this species is a mosaic that consists of the main 'contact A', as well as a second contact (B) involving a recent geographical isolate of the S lineage (iS hereafter) within the range of the N lineage (Fig. 1b). Here we test whether speciation by reinforcement has occurred between the two lineages at either contact A or contact B, and, if so, whether this has incidentally driven allopatric speciation between the iS population and the genetically similar main range of the S lineage.

Genetic analysis revealed that hybridization between the lineages occurs in the field, although it is geographically broader (6.0 km versus 0.6 km), and significantly more frequent (3.1–6.8% hybrids versus 0–1.4%; likelihood ratio test, $LI = 8.46$, $P = 0.015$), at contact A than at contact B. For the process of reinforcement to operate, there must be selection against hybridization¹⁴. We tested this by performing experimental crosses between the N and S/iS lineages. Viability of hybrid crosses was highly asymmetric (likelihood ratio test, $LI = 20.73$, $P < 0.001$): all crosses involving an S or iS female with an N male failed at the early tadpole stage, whereas all the reciprocal crosses succeeded to termination of the experiment at late tadpole stage or metamorphosis, albeit with significantly slower development than the control crosses within the N lineage (univariate analysis of variance (ANOVA), $F_{1,10} = 9.526$, $P = 0.012$). The asymmetry in viability is supported by the genetic analysis of the contact zones, which revealed that none of the hybrids carried S mitochondrial DNA (mtDNA). There is therefore strong selection against hybridization involving S or iS females and potentially mild selection against hybrids from N female crosses.

We assessed character displacement in morphology and mating call in the contact zone. A pattern of character displacement is a potential signature of reinforcement, although it could also be caused by other factors^{4,13}. Morphology is not the direct determinant of mate choice in these frogs; however, it was assessed to test whether any divergence in this trait is linked to the mate choice trait (male call)^{14,15} or to ecological divergence. As found for other species from the Wet Tropics with divergent phylogeographic lineages¹⁶, there was no difference between N and S body size or shape away from the contact for either sex (Supplementary Information). Further, across the contact zone, there was no difference between N, S and iS for female

size and male shape (Supplementary Information). The only morphological difference detected was a 20% reduction in mean body size of iS males in contact B (Fig. 2). Like most frogs¹⁵, mate choice in *L. genimaculata* is mediated primarily by female choice of a single trait—male call. This makes frogs one of the best groups for studying reinforcement¹⁴. An analysis of field recordings across the contact region revealed that the N, S and iS male calls all differ significantly from one another (Fig. 3). The iS males have diverged in size and call from all other S males, whereas the S males at contact A do not differ from those outside the contact region (Supplementary Information and Supplementary Figs 3, 4).

If reinforcement is operating in this system, we can make two predictions about the impact of call divergence on mate choice. First, southern females should choose correctly more often than northern females because the cost of hybridization is greater for them, and second, the greater divergence in call of iS males should result in stronger assortative mating at contact B than at contact A. Consistent with these predictions, female two-choice trials revealed the following: first, females of the southern lineage (iS and S) were significantly better at choosing their own lineage than northern females (Fig. 4), and second, there was significantly stronger premating isolation at contact B than at contact A, to the point of highly significant positive assortative mating between the lineages at contact B (Fig. 4). The genetic results support these patterns in that there are significantly fewer hybrids at contact B (0–1.4%) than at contact A (3.1–6.8%), and none with S mtDNA. The almost complete reproductive isolation at contact B leads us to conclude that the two lineages have speciated at contact B but not at contact A.

Above, we have satisfied four of Howard's⁴ five criteria for demonstrating reinforcement, and furthermore we have shown that this process has resulted in speciation: hybridization occurs in the field; there is strong postzygotic selection against hybrids; mate choice in this system is determined primarily by a single genetically determined character, frog call¹⁵; and divergence in this character at contact B has enhanced premating isolation between the two lineages to the point of significant reproductive isolation. The final requirement is to disentangle the effects of ecology on trait divergence^{4,13}. Divergence in body size and call of the iS males does not seem to relate to ecological factors for the following reasons: first, females do not show size divergence across the contact region, and there is no divergence in male shape, as might be expected for ecologically driven divergence; and second, the distributions of N, S and iS do not show any relationship to habitat characteristics (for vegetation,

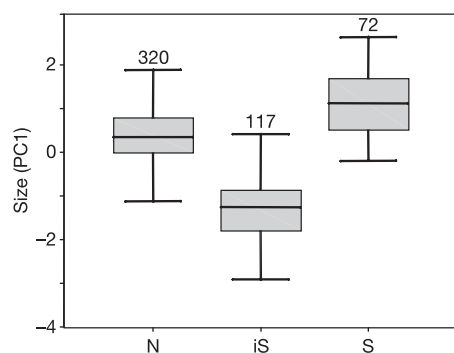


Figure 2 | Divergence in male body size across the mosaic contact zone.

The box plots show the median, 25th and 75th quartiles, and minimum and maximum data of the first principal component (PC1). Numbers shown above the box plots indicate the number of individuals. PC1 represents body size and explains 90% of the variation between N, iS and S. Male size differs significantly (ANOVA on PC1, $F_{2,24} = 35.42$, $P < 0.001$) because iS males are significantly smaller than both N males (contrast, $F_{1,16} = 20.63$, $P < 0.001$) and S males (contrast, $F_{1,16} = 18.58$, $P < 0.001$).

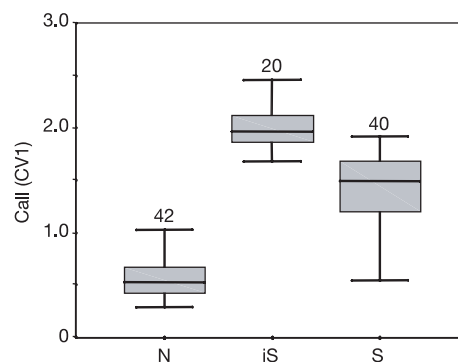


Figure 3 | Divergence in call across the mosaic contact zone.

The box plots show the median, 25th and 75th quartiles, and minimum and maximum data of the first canonical variate (CV1). Numbers shown above the box plots indicate the number of individuals. CV1 explains 93.8% of call variation between N, iS and S. Call differs significantly between the groups (MANOVA on three call characters, $F_{6,12} = 11.28$, $P < 0.001$), with the S call differing significantly from the N call (contrast, $F_{3,6} = 12.80$, $P = 0.005$), and the iS call differing significantly from both the N (contrast, $F_{3,6} = 34.67$, $P < 0.001$) and the S calls (contrast, $F_{3,6} = 9.29$, $P = 0.011$).

LI = 7.77, $P = 0.133$; for stream substrate, LI = 1.49, $P = 0.837$; for stream flow, LI = 0.41, $P = 0.888$).

The call divergence of iS, resulting in speciation from N at contact B, has incidentally led to significant divergence of the iS call from that of the remainder of the southern lineage (Fig. 3 and Supplementary Fig. 4). To test whether this call divergence has resulted in premating isolation, mate choice trials were conducted in which iS and S females could choose between an iS and an S male call. There was complete premating isolation—in all experiments both iS and S females chose the male call of their own group (likelihood ratio test, LI = 19.12, $P < 0.001$). This allopatric speciation within the southern lineage seems to have occurred rapidly because there is very low sequence divergence between iS and S and they share haplotypes (Fig. 1). This demonstrates a twofold role of natural selection in speciation: natural selection acting directly against hybridization in a contact zone can also indirectly drive rapid allopatric speciation. This significantly extends the role of reinforcement from previous work^{5,6} by showing, unconfounded by ecology and intervening populations, speciation of a contact zone population, both from another lineage and from its own lineage, as a result of reinforcement of a mate choice trait. Reinforcement has been viewed to have a role only in the final stages of speciation between already differentiated lineages^{1,17}, whereas the present results show the potential for reinforcement to be the sole cause of speciation.

Questions arising from this research include the following. First, why are iS males so small, given the lack of morphological divergence in the iS females? We propose that the reduction in male size might have been driven by sexual selection on iS male call. Because of a strong relationship between call divergence and body size in iS (Supplementary Information, and Supplementary Fig. 5), the more divergent the iS male call chosen by the females is from the other groups, the smaller the selected male is. Second, given that the postmating penalty for hybridization by the southern lineage is the same at both contacts, why has reinforcement been effective at contact B but not at contact A? The primary difference between contacts A and B is the range size of the southern lineage parental population—at contact A it is many times greater than the contact zone, whereas at contact B it is of smaller width than the contact zone. Our data are consistent with the suggestion^{4,18,19} that if the parental

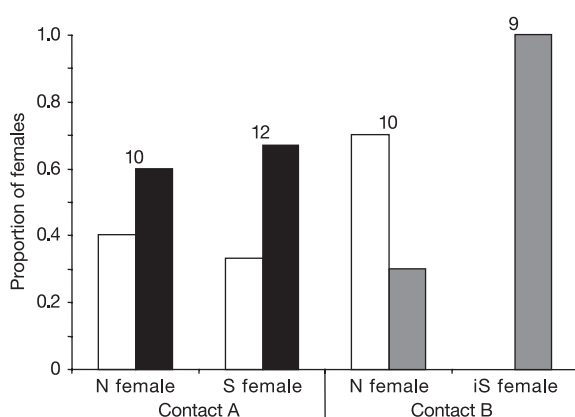


Figure 4 | Female choice of male calls at contacts A and B. Females of the southern lineage (iS and S) displayed significantly better choice, corrected for the effect of contact, than N females (generalized linear model, change in deviance $\Delta_{\text{dev}} = 4.33$, degrees of freedom, d.f. = 1, $P = 0.037$). Further, premating isolation, corrected for lineage, was significantly stronger at contact B than at contact A (generalized linear model, $\Delta_{\text{dev}} = 5.42$, d.f. = 1, $P = 0.020$). There is significant premating isolation between the northern and southern lineages at contact B (likelihood ratio test, LI = 12.79, $P = 0.003$) but not at contact A (likelihood ratio test, LI = 0.105, $P = 1.000$). White columns, N male; black columns, S male; grey columns, iS male. Numbers shown indicate the number of females.

population area is sufficiently small relative to both the area of overlap and the dispersal distance (not measured in this study), the diluting effects of gene flow from the parental population would be limited^{20–22}, and enough of the population may be exposed to maladaptive hybridization that a reinforced trait could spread throughout^{1,2,23}, resulting in speciation. The prevalence of such speciation in nature may be limited by the likelihood of extinction of the small population on initial contact¹⁷.

METHODS

For full details of all methods see Supplementary Information.

Sampling and genetic analyses. Sampling sites are presented in Supplementary Figs 1 and 2. Sampling at contacts A and B consisted of stream transects and scattered sites covering the breadth of the zone. Genetic material was taken from toe-pads. The tree was constructed from 527 base pairs of *COI* mtDNA (primers Cox and Coy¹¹) from 141 individuals (58 N, 66 S and 17 iS) from across the range, by using neighbour-joining with Kimura two-parameter distance estimates. *COI* (sequencing and restriction-fragment-length polymorphism test) was used to screen 787 individuals from 68 sites across 40 drainages. The contact zone was characterized by using *COI* and two nuclear markers.

Two near-diagnostic nuclear loci, LITGEN02 and CRYBA²⁴, were screened across 619 individuals: 63 N and 80 S (6 drainages in each) from outside of the contact region, and 248 (13 sites, 4 drainages) and 228 (10 sites, 4 drainages) from contacts A and B, respectively. The status of individuals was inferred by using a combination of mtDNA and bayesian inference (using 'NewHybrid'²⁵) of nuclear genotypes. We present percentage hybrids from both a stringent criterion, P (mismatch of nuclear genotype to mitochondrial genotype) > 0.9 , and a more relaxed criterion, $0.7 > P > 0.5$, for identifying potential hybrids. Both criteria gave similar results.

Statistical analyses: general comments. All analyses of morphology and call were nested by drainage. The planned, non-orthogonal contrasts—N versus S, N versus iS, and S versus iS—were used for the analyses of morphology and calls. Significance values were compared with sequential Bonferroni values²⁶. All two-way contingency tables were analysed as tests of independence²⁶. Given the relatively small expected frequencies in some cells, exact tests (likelihood ratio tests) were performed^{26,27}.

Experimental crosses. Males and gravid females were collected from the field and paired in the lab until eggs were laid. Once the hatchlings of a clutch reached stage 23 (ref. 28), two replicates of ten tadpoles each were placed in random positions beneath ultraviolet lamps and raised on lettuce. The difference in viability of inter-lineage crosses involving N ($n = 8$) and S/iS ($n = 6/1$) females was analysed with a likelihood ratio test. An ANOVA was used to test the difference in development rate (time until hindlimb buds, stage 26 (ref. 28)) between the N female hybrid clutches ($n = 8$) and the N lineage control clutches ($n = 4$).

Morphological analyses. Measurements taken were snout-to-vent length (SVL), tarsus length, head width, and weight. A total of 924 mature males (34 drainages) and 153 mature females (23 drainages) were measured. To remove confounding effects of altitude, analyses of covariance (ANCOVAs) were performed on each of the morphological traits, followed by linear regressions from which the unstandardized residuals were taken. Principal-component analyses (PCAs) were conducted because the traits were all highly correlated. The first principal component (PC1) was used as the measure of body size in nested ANOVAs. The remaining PCs were used in nested multivariate analyses of variance (MANOVAs) to assess differences in body shape. To maximize drainage coverage for the analyses of size divergence in females across the contact region and males across the S lineage, only SVL was used in nested ANOVAs; shape was not assessed.

Call analyses. Call duration, dominant frequency and note rate were measured in four calls in each of 132 males (19 drainages). To remove confounding effects of air temperature, ANCOVAs were performed on the three call characters, followed by linear regressions from which the unstandardized residuals were taken. Nested MANOVAs were run on the three call characters, followed by multivariate contrasts. Nested canonical variates (CVs) were calculated for presentation and analysis of call divergence against body size.

Mate choice. Gravid females were presented with two field-recorded calls in a laboratory mate-choice chamber. Females from contact A ($n = 22$) were presented with an N and an S call, and those from contact B ($n = 19$) were presented with an N and an iS call. After a 5-min listening period, a 10-min trial was conducted. The trials to assess premating isolation between S ($n = 8$ females) and iS ($n = 6$ females) were performed in the same manner except that some of the females and male calls came from outside the contact region.

The difference in mate choice between the southern (S and iS) and northern (N) lineages, and the difference in premating isolation at contacts A and B, were compared by using a generalized linear model with a binomial error specified. The level of premating isolation at each contact, and premating isolation between S and iS were analysed with likelihood ratio tests.

Habitat association. Three broad habitat characteristics (stream-side vegetation, stream substrate and stream flow) were classified into categories at each site ($n = 46$) in the contact region. The association between N, S and iS and each of the three habitat characteristics was assessed by using a likelihood ratio test.

Received 10 May; accepted 11 July 2005.

- Dobzhansky, T. *Genetics and the Origin of Species* 3rd edn (Columbia Univ. Press, New York, 1951).
- Mayr, E. *Animal Species and Evolution* 548–555 (Belknap Press, Cambridge, Massachusetts, 1963).
- Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).
- Howard, D. J. in *Hybrid Zones and the Evolutionary Process* (ed. Harrison, R. G.) 46–69 (Oxford Univ. Press, New York, 1993).
- Littlejohn, M. J. & Loftus-Hills, J. J. An experimental evaluation of premating isolation in the *Hyla ewingi* complex (Anura: Hylidae). *Evolution* **22**, 659–663 (1968).
- Zouros, E. & d'Entremont, C. J. Sexual isolation among populations of *Drosophila mojavensis*: response to pressure from a related species. *Evolution* **34**, 421–430 (1980).
- Servedio, M. R. & Noor, M. A. The role of reinforcement in speciation: theory and data. *Annu. Rev. Ecol. Syst.* **34**, 339–364 (2003).
- Butlin, R. K. Reinforcement: an idea evolving. *Trends Ecol. Evol.* **10**, 433–434 (1995).
- Butlin, R. K. Mystery of mysteries no longer? *Evolution* **58**, 243–245 (2004).
- Butlin, R. K. & Tregenza, T. Evolutionary biology: is speciation no accident? *Nature* **387**, 551–552 (1997).
- Schneider, C. J., Cunningham, M. & Moritz, C. Comparative phylogeography and the history of endemic vertebrates in the Wet Tropics rainforest of Australia. *Mol. Ecol.* **7**, 487–498 (1998).
- Phillips, B. L., Baird, S. J. E. & Moritz, C. When vicars meet: a narrow contact zone between morphologically cryptic phylogeographic lineages of the rainforest skink, *Carlia rubrigularis*. *Evolution* **58**, 1536–1549 (2004).
- Noor, M. A. Reinforcement and other consequences of sympatry. *Heredity* **83**, 503–508 (1999).
- Blair, W. F. Isolating mechanisms and interspecies interactions in anuran amphibians. *Q. Rev. Biol.* **39**, 333–344 (1964).
- Gerhardt, H. C. & Huber, F. *Acoustic Communication in Insects and Anurans* (Univ. Chicago Press, Chicago, 2002).
- Schneider, C. J. & Moritz, C. Rainforest refugia and evolution in Australia's Wet Tropics. *Proc. R. Soc. Lond. B* **266**, 191–196 (1999).
- Liou, L. W. & Price, T. D. Speciation by reinforcement of premating isolation. *Evolution* **48**, 1451–1459 (1994).
- Littlejohn, M. J. in *Evolution and Speciation: Essays in Honor of M. J. D. White* (eds Atchley, W. R. & Woodruff, D. S.) 298–334 (Cambridge Univ. Press, Cambridge, 1981).
- Barton, N. H. & Hewitt, G. M. Adaptation, speciation and hybrid zones. *Nature* **341**, 497–503 (1989).
- Bigelow, R. S. Hybrid zones and reproductive isolation. *Evolution* **19**, 449–458 (1965).
- Sanderson, N. Can gene flow prevent reinforcement? *Evolution* **43**, 1223–1235 (1989).
- Cain, M. L., Andreasen, V. & Howard, D. J. Reinforcing selection is effective under a relatively broad set of conditions in a mosaic hybrid zone. *Evolution* **53**, 1343–1353 (1999).
- Moore, J. A. in *The Species Problem* (ed. Mayr, E.) 325–338 (American Association for the Advancement of Science, Washington DC, 1957).
- Dolman, G. & Phillips, B. Single copy nuclear DNA markers characterized for comparative phylogeography in Australian wet tropics rainforest skinks. *Mol. Ecol. Notes* **4**, 185–187 (2004).
- Anderson, E. C. & Thompson, E. A. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229 (2002).
- Sokal, R. R. & Rohlf, F. J. *Biometry: the Principles and Practice of Statistics in Biological Research* 724–740 (W. H. Freeman, New York, 1995).
- Agresti, A. *Categorical Data Analysis* 239–249 (Wiley, New York, 1990).
- Gosner, K. A simplified table for staging anuran embryos and larvae with notes on identification. *Herpetologica* **16**, 183–190 (1960).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Phillips, J. MacKenzie, M. Tonione, J. Gardiner, M. Blows, J. Austin, M. Cunningham, H. McCallum, G. Dolman, S. Williams, H. Rundle, S. Chenoweth, A. Freeman, F. J. Rohlf and D. Wake. We are also grateful to B. Phillips and M. Cunningham for locating the contact zone. Supported by the National Science Foundation (C.M.), an Australian Postgraduate Award (C.J.H.), a University of Queensland Graduate School Research Travel Award (C.J.H.), the Cooperative Research Centre for Tropical Rainforest Ecology and Management (C.J.H.) and Queensland Parks and Wildlife Service.

Author Information Sequences are deposited in the EMBL database under the following accession numbers: AF304205–AF304229 (ref. 11) and AJ872186–AJ872201. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.J.H. (c.hoskin@sib.uq.edu.au).

Chimpanzees are indifferent to the welfare of unrelated group members

Joan B. Silk¹, Sarah F. Brosnan^{2,3}, Jennifer Vonk⁴, Joseph Henrich², Daniel J. Povinelli⁴, Amanda S. Richardson³, Susan P. Lambeth³, Jenny Mascaro³ & Steven J. Schapiro³

Humans are an unusually prosocial species—we vote, give blood, recycle, give tithes and punish violators of social norms. Experimental evidence indicates that people willingly incur costs to help strangers in anonymous one-shot interactions^{1,2}, and that altruistic behaviour is motivated, at least in part, by empathy and concern for the welfare of others (hereafter referred to as other-regarding preferences)^{1–3}. In contrast, cooperative behaviour in non-human primates is mainly limited to kin and reciprocating partners, and is virtually never extended to unfamiliar individuals⁴. Here we present experimental tests of the existence of other-regarding preferences in non-human primates, and show that chimpanzees (*Pan troglodytes*) do not take advantage of opportunities to deliver benefits to familiar individuals at no material cost to themselves, suggesting that chimpanzee behaviour is not motivated by other-regarding preferences. Chimpanzees are among the primates most likely to demonstrate prosocial behaviours. They participate in a variety of collective activities, including territorial patrols, coalitionary aggression, cooperative hunting, food sharing and joint mate guarding^{5–12}. Consolation of victims of aggression¹³ and anecdotal accounts of solicitous treatment of injured individuals suggest that chimpanzees may feel empathy^{14,15}. Chimpanzees sometimes reject exchanges in which they receive less valuable rewards than others, which may be one element of a ‘sense of fairness’, but there is no evidence that they are averse to interactions in which they benefit more than others^{16–18}.

We conducted two experiments to assess the existence of other-regarding preferences in two different populations of chimpanzees. In both experiments, subjects were able to deliver benefits to others at no cost to themselves. Subjects were presented with an apparatus that gave them a choice between two alternatives. If the subject (hereafter referred to as the actor) chose option 1, the actor obtained a food reward and another chimpanzee simultaneously received an identical reward (hereafter referred to as the ‘1/1 option’). If the actor chose option 2, the actor obtained the same size and type of food reward, but no food reward was delivered to the other chimpanzee (the ‘1/0 option’). As a control, actors were presented with exactly the same reward options when there was no other chimpanzee present.

If chimpanzees have other-regarding preferences, they will choose the 1/1 option more often when another chimpanzee is present to receive the reward than when they are alone. If chimpanzees are indifferent to the welfare of others, the presence of a potential recipient will have no impact on their choices. (Alternatively, chimpanzees might have ‘antisocial’ preferences. If so, they would choose the 1/1 option less often when another chimpanzee is present than when they are alone.)

This experimental setup maximizes the likelihood of observing other-regarding behaviour in two ways. First, actors can provide benefits to others at no cost to themselves, so other-regarding sentiments do not compete with selfish motives to obtain rewards. Second, actors interact with familiar group members. Prosocial responses in this experiment might occur because chimpanzees favour those that they cooperate with outside the context of this experiment, even if they lack other-regarding sentiments. However, the absence of prosocial behaviour in this experimental situation would provide strong evidence for the lack of other-regarding sentiments.

Our experiments were conducted at two different study sites. In Louisiana, our subjects were seven unrelated adults that have been living together for at least 15 yr, and have participated regularly in cognitive and behavioural tests since they were 3–4 yr old¹⁹. In Texas, our subjects were drawn from six stable social groups originally formed in 1978. We tested 11 same-sex pairs of adults drawn from the same social groups. These animals were well socialized, but had no experience in cognitive or behavioural testing before the present study began.

The physical layout of the testing areas differed, so we designed different apparatuses for each site (see Supplementary Information). At both sites, subjects were clearly able to view each other, the testing apparatus, the baiting process, the distribution of food rewards, and the consumption of food rewards. After the actor selected one option (by pulling a rope or hose), both were able to obtain food rewards (if present) from the trays closest to them. All trials were videotaped and coded by independent observers. For all measures, high levels of inter-observer reliability were obtained (Cohen's kappa ≥ 0.98).

At both sites, the subjects were familiarized individually with the apparatus before testing. Pre-testing established that the chimpanzees understood how to operate the apparatus to obtain food rewards for themselves and that they were highly motivated to do so. Pre-testing also demonstrated that the chimpanzees understood that they would obtain food only from the near tray (Louisiana) or their own side (Texas).

In Louisiana, actors chose the 1/1 option on average 56% (s.d. = 8.2) of the time when they were alone, and 58% (s.d. = 6.3) of the time when another chimpanzee was present. None of the seven chimpanzees chose the 1/1 option significantly more often when another chimpanzee was present than when they were alone (Fig. 1). A power analysis indicates that sample sizes were large enough to detect even modest differences between the two conditions. If actors were 15% more likely to choose the 1/1 option when another chimpanzee was present than when they were alone, the chance of

¹Department of Anthropology, University of California, Los Angeles, California 90095, USA. ²Department of Anthropology, Emory University, Atlanta, Georgia 30322, USA.

³Michale E. Keeling Center for Comparative Medicine and Research, The University of Texas M. D. Anderson Cancer Center, Bastrop, Texas 78602, USA. ⁴Cognitive Evolution Group, University of Louisiana at Lafayette, New Iberia, Louisiana 70560, USA.

failing to detect a difference of this magnitude in any of the subjects is just 3.5%.

Multivariate analyses of the Louisiana data set indicate that the likelihood of choosing the 1/1 option was more strongly influenced by the position of the 1/1 option than by the presence of another chimpanzee (Table 1). The odds ratios indicate that actors were 95% more likely to choose the 1/1 option when it was presented on the right side than on the left side of the apparatus ($P < 0.001$). In contrast, the presence of another chimpanzee increased the chance of choosing the 1/1 option by only 11% ($P < 0.580$). The chimpanzees were more likely to choose the 1/1 option as the experiment progressed, but this increase was not influenced by the presence of another chimpanzee, and the interaction between trial block and condition (partner present/absent) was nonsignificant (see Supplementary Table 6).

In Texas, not all subjects responded on every trial and not all subjects completed all sessions. Therefore, we computed the proportion of 1/1 choices by dividing the number of 1/1 choices by the total number of responses. On average, actors chose the 1/1 option 48% of the time (s.d. = 7.2) when they were alone and 48% of the time (s.d. = 16.7) when another chimpanzee was present. None of the actors chose the 1/1 option significantly more often when another chimpanzee was present than when they were alone (Fig. 2). The chance of failing to detect a 15% difference between the partner present and partner absent conditions in any of these subjects is just 0.2%. (When we combine subjects from both sites, the chance that we would fail to detect at least one prosocial chimpanzee is only 0.006%.)

Again, the likelihood of choosing the 1/1 option was more strongly affected by the actors position than by the presence of another chimpanzee in the adjoining enclosure. The probability of choosing the 1/1 option was 56% greater when the 1/1 option was positioned in the upper tray than the lower tray ($P < 0.001$), whereas the presence of a potential recipient increased the probability of choosing the 1/1 option by only 11% ($P < 0.40$). There was no significant change in the likelihood of choosing the 1/1 option across sessions.

In trials in which only one tray was baited on the actor's side (1/0 or 1/1), actors chose the option that provided food to themselves on 92% (105 out of 114; 94 no response) of all trials when they were alone and 94% (106 out of 113; 79 no response) of all trials when another chimpanzee was present. Thus, actors were attending closely

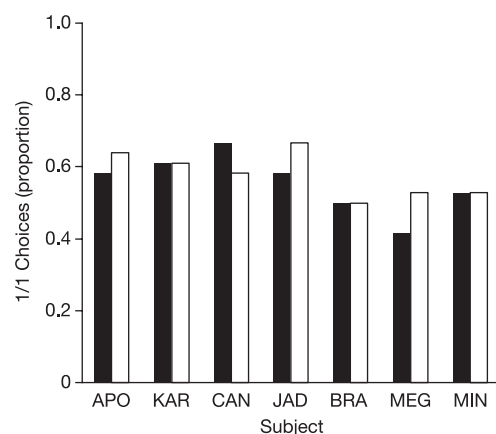


Figure 1 | Proportion of 1/1 choices made by Louisiana subjects. Black bars represent sessions in which the subject was alone. White bars represent sessions in which another chimpanzee was present in the opposite enclosure. All subjects completed 36 single trial sessions for each condition. None of the subjects chose the 1/1 option significantly more often when another chimpanzee was present than when they were alone (Fisher's exact test: APO, $P = 0.809$; KAR, $P = 1.000$; CAN, $P = 0.627$; JAD, $P = 0.479$; BRA, $P = 1.000$; MEG, $P = 0.479$; MIN, $P = 1.000$).

Table 1 | Factors that influence the likelihood of choosing the 1/1 option

Parameter*	Estimate	Standard error	t-ratio	P-value	Odds ratio	95% bounds	
						Upper	Lower
Louisiana							
Condition	0.102	0.185	0.553	0.580	1.107	1.590	0.771
Position	0.666	0.185	3.602	<0.001	1.946	2.796	1.355
Trial block	0.117	0.054	2.162	0.031	1.124	1.251	1.011
Texas							
Condition	0.108	0.128	0.844	0.398	1.114	1.431	0.867
Position	0.445	0.125	3.572	<0.001	1.561	1.993	1.223
Session	-0.006	0.024	-0.250	0.803	0.994	1.041	0.949

* The two logistic regression models include dummy variables for individuals as predictors, in addition to the predictor variables shown (see Supplementary Information for details). Variables are coded so that the odds ratios would exceed 1 for condition if actors were more likely to choose the 1/1 option when a potential recipient was present than when they were alone. Position was coded so that odds ratios >1 indicate that actors were more likely to choose the 1/1 option when it was positioned on the right side (Louisiana) or the top tray (Texas) of the apparatus. Finally, odds ratios for trial block (in Louisiana) and session (Texas) would exceed 1 if actors were more likely to choose the 1/1 option as the experiment progressed.

to the distribution of payoffs for themselves, not picking options at random.

It is possible that the chimpanzees in our experiments understood how to obtain food for themselves but not that they were responsible for delivering rewards to the chimpanzee in the adjoining enclosure. Several factors mitigate against this alternative. Chimpanzees are adept at manipulating push/pull apparatuses like the one used in Louisiana, and understand the necessity of their actions in generating contingent effects²⁰. The bar-pull apparatus used in Texas was similar to apparatuses that have been widely used in studies of contingent cooperation and mutualism in non-human primates^{21–24}. Actors and potential recipients could see and hear each other at both sites, and actors had many opportunities to watch recipients consume food rewards after they chosen the 1/1 option. In Louisiana, all subjects participated in the experiment as both actors and potential recipients. Potential recipients sometimes displayed begging gestures, suggesting that they had some understanding of the actor's role in delivering rewards to them. Finally, results of a second set of experiments using an entirely different apparatus (in Louisiana) and slightly different protocol (in Texas) generated results very similar to the ones reported here.

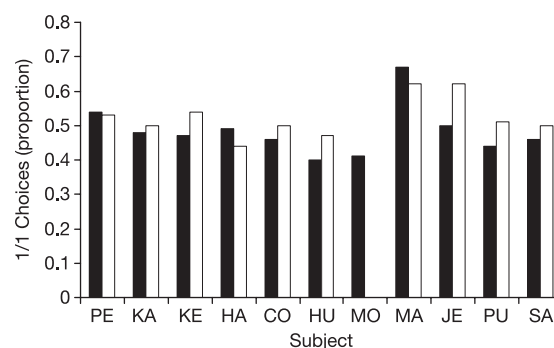


Figure 2 | Proportion of 1/1 choices made by Texas subjects. Values were obtained by dividing the number of 1/1 choices by the total number of 1/1 and 1/0 responses. Conventions as in Fig. 1. The first five subjects completed all sessions and responded on 69–99% of all trials. The next three subjects completed all sessions, but responded on only 18–27% of trials. MO never chose the 1/1 option when a potential recipient was present. The last three subjects did not complete all sessions, but responded on 71–88% of all trials that they began. None of the subjects chose the 1/1 option significantly more often when another chimpanzee was present than when they were alone (Fisher's exact test: PE, $P = 1.00$; KA, $P = 0.868$; KE, $P = 0.491$; HA, $P = 0.596$; CO, $P = 0.704$; HU, $P = 1.000$; MO, $P = 0.271$; MA, $P = 1.000$; JE, $P = 0.237$; PU, $P = 0.539$; SA, $P = 0.823$).

In these experiments, chimpanzees were clearly motivated to obtain rewards for themselves but not to provide rewards for other group members. None of the 18 chimpanzees that we tested was more likely to choose the 1/1 option when another chimpanzee was present than when they were alone. Our findings are strengthened by the fact that chimpanzees from both populations behaved in very similar ways, despite having different life histories and experimental experience.

These results complement observational and experimental studies that indicate that chimpanzees cooperate mainly with kin and reciprocating partners^{9–16} and show no aversion to inequitable exchanges that benefit themselves^{21,22}. The absence of other-regarding preferences in chimpanzees may indicate that such preferences are a derived property of the human species^{25,26} tied to sophisticated capacities for cultural learning, theory of mind, perspective taking and moral judgement^{27,28}. Alternatively, other-regarding preferences might be found in other species that rely more heavily on cooperative strategies than chimpanzees do, such as cooperatively breeding mammals²⁹. Further work on other species will help to clarify the socioecological conditions and cognitive requirements associated with the evolution of other-regarding preferences.

METHODS

Experimental procedure in Louisiana. The two chimpanzees faced each other in opposing enclosures. The testing apparatus spanned the width of an enclosure that adjoined two other enclosures (Supplementary Fig. 1). The apparatus consisted of two expandable arms, each connected to two horizontally aligned food trays. Each trial began with all four food trays collapsed towards the centre of the apparatus. When a handle was pulled, the closest tray on that arm moved to within the actor's reach and the other tray on the same arm extended in the opposite direction to within reach of the potential recipient (when present).

During testing each subject was paired with each of the other group members six times in random order in single trial sessions. Trials were alternated between partner-present and partner-absent sessions. The left/right position of the 1/1 option was counterbalanced within blocks of six trials on the partner-absent trials, and counterbalanced within pairings on the partner-present trials.

Experimental procedure in Texas. The chimpanzees were tested in the indoor runs of their home enclosure. The chimpanzees were put in adjoining cages separated by a wire mesh divider, and the testing apparatus was placed in front of them. The apparatus was placed between the two enclosures and consisted of two clear Lexan trays arrayed vertically (Supplementary Fig. 2). Each tray had a metal bar attached to the tray. A hose was attached to the bar on each tray; when the hose was pulled in, the bar swept forward across the tray and brought the food reward (if present) within the actor's reach. The actor and the recipient were able to reach the tray and remove rewards on their own side when the bar was pulled forward. When one bar was pulled forward, the other tray locked into position and the hose attached to it was retracted.

We conducted 10 sessions per pair, alternating between partner-present and partner-absent sessions. Within each session, 16 trials of the test configuration (1/1 versus 1/0) were conducted. The top/bottom position of the 1/1 option was counterbalanced across trials within sessions. Actors did not exchange roles or partners over the course of the experiment.

To reduce the likelihood that actors would stop attending to the distribution of rewards across trials within sessions, four additional 'attention trials' were randomly interspersed with the test configuration. In these trials, one option was baited only on the actor's side (1/0) and the other was baited only on the other side (0/1). If subjects are attending to the payoffs in these interspersed trials, they will choose the 1/0 option.

Received 6 July; accepted 20 September 2005.

1. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
2. Henrich, J. et al. (eds) *Foundations of Human Sociality* (Oxford Univ. Press, Oxford, 2004).
3. Penner, L. A., Dovidio, J. F., Piliavin, J. A. & Schroeder, D. A. Prosocial behaviour: multilevel perspectives. *Annu. Rev. Psychol.* **56**, 365–392 (2005).
4. Silk, J. B. in *Moral Sentiments and Material Interests: On the Foundations of*

Cooperation in Economic Life (eds Gintis, H., Bowles, S., Boyd, R. & Fehr, E.) 43–74 (MIT Press, Cambridge, 2005).

5. Mitani, J. & Watts, D. P. Why do chimpanzees hunt and share meat? *Anim. Behav.* **61**, 915–924 (2001).
6. Mitani, J., Merriwether, D. A. & Zhang, C. Male affiliation, cooperation, and kinship in wild chimpanzees. *Anim. Behav.* **59**, 885–893 (2000).
7. Watts, D. P. Coalitionary mate guarding by male chimpanzees at Ngogo, Kibale National Park, Uganda. *Behav. Ecol. Sociobiol.* **44**, 43–55 (1998).
8. Watts, D. P. Reciprocity and interchange in the social relationships of wild male chimpanzees. *Behaviour* **139**, 343–370 (2002).
9. Watts, D. P. & Mitani, J. Boundary patrols and intergroup encounters in wild chimpanzees. *Behaviour* **138**, 299–327 (2001).
10. Goodall, J. *The Chimpanzees of Gombe* (Harvard Univ. Press, Cambridge, 1986).
11. Boesch, C. & Boesch-Achermann, H. *The Chimpanzees of the Tai Forest* (Oxford Univ. Press, Oxford, 2002).
12. Hohmann, G., Gerloff, U., Tautz, D. & Fruth, B. Social bonds and genetic ties: kinship, association, and affiliation in a community of bonobos (*Pan paniscus*). *Behaviour* **136**, 1219–1235 (1999).
13. de Waal, F. B. M. & Aureli, F. in *Reaching into Thought* (eds Russon, A. E., Bard, K. A. & Parker, S. T.) 80–110 (Cambridge Univ. Press, Cambridge, 1996).
14. Flack, J. C. & de Waal, F. B. M. 'Any animal whatever': Darwinian building blocks of morality in monkeys and apes. *J. Consciousness Stud.* **7**, 1–29 (2000).
15. Preston, S. D. & de Waal, F. B. M. Empathy: its ultimate and proximate bases. *Behav. Brain Sci.* **25**, 1–72 (2002).
16. Brosnan, S. F. & de Waal, F. B. M. Monkeys reject unequal pay. *Nature* **425**, 297–299 (2003).
17. Brosnan, S. F., Schiff, H. C. & de Waal, F. B. M. Tolerance for inequity may increase with social closeness in chimpanzees. *Proc. R. Soc. Lond. B* **272**, 253–258 (2005).
18. Henrich, J. Inequity aversion in capuchins? *Nature* **428**, 139 (2004).
19. Povinelli, D. J. in *Developing Theories of Intention: Social Understanding and Self-Control* (eds Zelazo, P., Astington, J. & Olson, D.) 195–225 (L. Erlbaum, Hillsdale, 1999).
20. Povinelli, D. J. *Folk Physics for Apes* (Oxford Univ. Press, Oxford, 2003).
21. Crawford, M. The cooperative solving of problems by young chimpanzees. *Comp. Psych. Monogr.* **14**, 1–88 (1937).
22. Hauser, M. D., Chen, M. K., Chen, F. & Chuang, E. Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back. *Proc. R. Soc. Lond. B* **270**, 2363–2370 (2003).
23. De Waal, F. B. M. & Berger, M. Payment for labour in monkeys. *Nature* **404**, 563 (2000).
24. Cronin, K. A., Kurian, A. V. & Snowdon, C. T. Cooperative problem solving in a cooperatively breeding primate (*Saguinus oedipus*). *Anim. Behav.* **69**, 133–142 (2005).
25. Richerson, P. J. & Boyd, R. *Not by Genes Alone: How Culture Transformed Human Evolution* (Univ. Chicago Press, Chicago, 2004).
26. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535 (2003).
27. Greene, J. & Haidt, J. How (and where) does moral judgement work? *Trends Cog. Sci.* **16**, 517–523 (2002).
28. Povinelli, D. P. & Godfrey, L. R. in *Evolutionary Ethics* (eds Nitecki, M. H. & Nitecki, D. V.) 277–324 (SUNY, Albany, 1993).
29. Clutton-Brock, T. H. Behavioral ecology—breeding together: kin selection and mutualism in cooperative vertebrates. *Science* **296**, 69–72 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a grant from the MacArthur Foundation Preferences Network to J.B.S., J.H. and D.J.P., and a grant from the James S. McDonnell Foundation to D.J.P. A James S. McDonnell Foundation Centennial Fellowship was awarded to D.J.P. S.F.B. was supported by an NIH/NIGMS IRACDA grant awarded to Emory University. The Texas colony was supported by NIH/NCRR. We are grateful to R. McElreath for advice about statistical procedures.

Author Contributions The original idea for the experiments was conceived by J.B.S. The experimental protocol used in Louisiana was developed by D.J.P., J.V. and J.B.S. J.V. supervised experiments, coding and data tabulation in Louisiana. The experimental protocol used in Texas was developed by S.F.B., J.H., J.V., D.J.P. and J.B.S. S.F.B. supervised experiments, coding and data tabulation in Texas; A.S.R. and J.M. assisted in data collection in Texas. S.P.L. and S.J.S. provided essential support for research at the Texas colony. J.H. conducted the statistical analyses. J.B.S. drafted the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.B.S. (jsilk@anthro.ucla.edu).

LETTERS

Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak

James A. Fraser^{1,2}, Steven S. Giles³, Emily C. Wenink¹, Scarlett G. Geunes-Boyer³, Jo Rae Wright³, Stephanie Diezmann^{1,4}, Andria Allen^{1,4}, Jason E. Stajich^{1,4}, Fred S. Dietrich^{1,4}, John R. Perfect⁵ & Joseph Heitman^{1,2,5,6}

Genealogy can illuminate the evolutionary path of important human pathogens. In some microbes, strict clonal reproduction predominates, as with the worldwide dissemination of *Mycobacterium leprae*, the cause of leprosy¹. In other pathogens, sexual reproduction yields clones with novel attributes, for example, enabling the efficient, oral transmission of the parasite *Toxoplasma gondii*². However, the roles of clonal or sexual propagation in the origins of many other microbial pathogen outbreaks remain unknown, like the recent fungal meningoencephalitis outbreak on Vancouver Island, Canada, caused by *Cryptococcus gattii*³. Here we show that the *C. gattii* outbreak isolates comprise two distinct genotypes. The majority of isolates are hypervirulent and have an identical genotype that is unique to the Pacific Northwest. A minority of the isolates are significantly less virulent and share an identical genotype with fertile isolates from an Australian recombining population. Genotypic analysis reveals evidence of sexual reproduction, in which the majority genotype is the predicted offspring. However, instead of the classic α - α sexual cycle, the majority outbreak clone appears to have descended from two α mating-type parents. Analysis of nuclear content revealed a diploid environmental isolate homozygous for the major genotype, an intermediate produced during same-sex mating. These studies demonstrate how cryptic same-sex reproduction can enable expansion of a human pathogen to a new geographical niche and contribute to the ongoing production of infectious spores. This has implications for the emergence of other microbial pathogens and inbreeding in host range expansion in the fungal and other kingdoms.

Cryptococcus neoformans is a haploid yeast, distributed worldwide in association with pigeon guano, and is the most common cause of fungal meningoencephalitis in immunosuppressed hosts⁴. In contrast, the less common sibling species *Cryptococcus gattii*⁵ is restricted to tropical and subtropical climes, inhabiting *Eucalyptus* trees and apparently infecting immunocompetent individuals. However, in 1999 a *C. gattii* outbreak emerged on Vancouver Island, Canada. Infection of otherwise healthy individuals and animals on the island is ongoing³, indicating an expanded geographical range for this pathogen.

The suspected infectious propagules of *Cryptococcus* are airborne spores^{4,6,7}, and although spores are produced during sexual reproduction, *Cryptococcus* mating has never been observed directly in nature⁸. Our previous studies have revealed that most Vancouver Island outbreak isolates are sexually fertile, but all are α mating type⁹. Although this precludes the traditional α - α sexual cycle⁸, Vancouver Island air samples contain particles of 1–2 μ m in diameter, a size consistent with spores¹⁰. One clue as to how spores might be produced in this unisexual population was suggested by studies

showing that in the laboratory, *C. neoformans* can undergo same-sex mating between two α mating-type partners¹¹. However, there is no evidence that this atypical sexual cycle occurs in nature.

To elucidate the origin of the Vancouver Island outbreak and determine whether sex had a role, a large-scale genealogical analysis was conducted with *C. gattii* isolates collected worldwide. Previous studies of this species by amplified fragment length polymorphism (AFLP) and randomly amplified polymorphic DNA (RAPD) analyses revealed four major molecular groups, designated VGI, VGII, VGIII and VGIV (ref. 12). Virtually all (>97%) of the Vancouver Island outbreak isolates belong to the VGII class^{10,12}. To establish robust genealogical relationships with greater precision, we used multilocus sequence typing (MLST) using nine polymorphic loci^{13,14}. Seven loci were from unlinked genomic regions present in all strains,

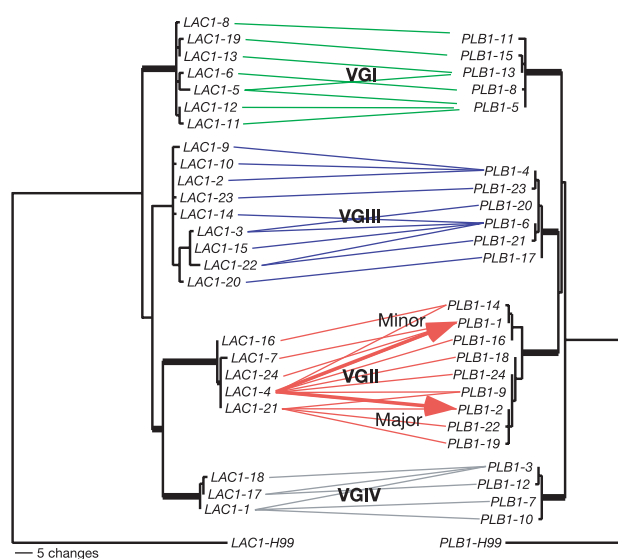


Figure 1 | Congruence of gene phylogeny in *C. gattii*. Maximum parsimony analyses of multiple loci from *C. gattii* supports the presence of four discrete molecular types (see also Supplementary Fig. 1 and Supplementary Table 1). Each allele shown is present in one or more isolates from the analysed population, with coloured lines showing combinations of the *LAC1* (encoding laccase) and *PLB1* (encoding phospholipase B) alleles in individual genomes. The Vancouver Island major and minor genotypes are indicated with red arrows. One thousand bootstrap replicates were sampled for the calculated statistical support. Thickened lines indicate branches with bootstrap values $\geq 70\%$. *C. neoformans* clinical strain H99 served as an outgroup.

¹Department of Molecular Genetics and Microbiology, ²Howard Hughes Medical Institute, ³Department of Cell Biology, ⁴Institute of Genome Sciences and Policy, ⁵Department of Medicine and ⁶Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA.

and two additional loci corresponded to the *SXI1 α* and *SXI2 α* mating-type-specific genes^{15,16}. Each strain contained only one *SXI* gene, molecularly defining mating type as **a** or α . We performed MLST analysis for 202 environmental and clinical isolates (77 VGI, 75 VGII, 26 VGIII, 24 VGIV) and the majority were α mating type (183 out of 202). Of those **a** strains available, fifteen belonged to class VGI, one to VGII, three to VGIII and none to VGIV. Together, the non-MAT MLST sequences totaled ~4,376 nucleotides per strain, from which 273 informative polymorphic sites were identified, representing 78 genotypes when the MAT-specific locus was included (Supplementary Table 1).

Concordant with preliminary RAPD analyses of population structure¹⁰, the Vancouver Island VGII class isolates fell into two discrete genotypes: a major form common in the environment and clinic, and a rarer, minor form represented by one clinical and several environmental samples. This demonstrates that there are two independent clonal populations present on Vancouver Island. Comparison of the two genotypes with worldwide isolates yielded three important insights. First, the major outbreak genotype is identical across all eight loci with the *C. gattii* type strain NIH444, which was isolated from a human sputum sample in Seattle thirty years ago, and with CBS7750, collected in San Francisco in 1992 from a *Eucalyptus camaldulensis* tree¹⁷. Second, the minor outbreak

genotype is identical across all eight loci with multiple isolates from Australia that belong to an unusual fertile, sexually recombining population^{18,19}—in contrast to the majority of Australian *C. gattii* isolates, which are sterile and non-recombining²⁰. Finally, the Vancouver Island major and minor genotypes are identical for three of the eight loci, and are therefore closely related to each other and more divergent from most other strains studied.

To further characterize the relationship between the Vancouver Island major and minor genotypes, and their relationship to a selection of similar strains worldwide, we analysed an additional 22 loci distributed throughout the genome (Supplementary Table 2). Close genetic relationships were observed for a number of the VGII strains selected (Supplementary Information). Most importantly, the Vancouver Island major genotype strain R265, the Seattle isolate NIH444 and the San Francisco isolate CBS7750 were identical across all 30 loci, indicating that the major genotype has been present in the Pacific Northwest for at least thirty years. An isolate with a similar genotype differing at only one locus was identified from Brazil (strain ICB107). The Vancouver Island minor genotype strain R272 and the fertile Australian isolate NT-13 were also identical across all 30 loci, indicating that the minor outbreak genotype might have originated in Australia. An isolate with a similar genotype differing at only one locus was identified from the Caribbean (strain 99/473). Comparison

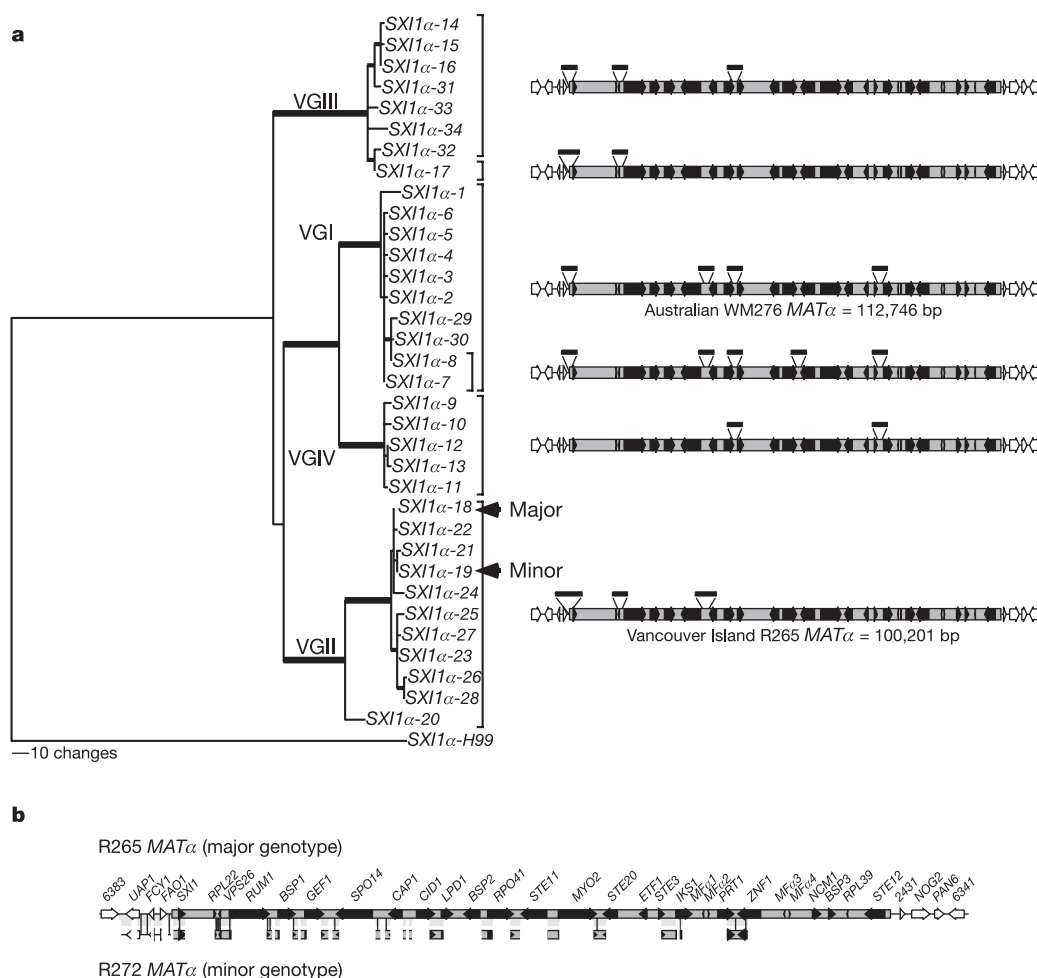


Figure 2 | MAT locus inheritance. **a**, Maximum parsimony analysis shows that the major and minor outbreak *SXI1 α* alleles are different. PCR-based fingerprinting of 16 intergenic regions within and flanking *MAT* revealed six different fingerprints defined by differences in intergenic length (represented by wedges); however, the two VGII outbreak genotype fingerprints were identical (Supplementary Table 3). Fingerprints were scaffolded on the WM276 *MAT* locus⁹, which is syntenic with the R265

MAT α allele. One thousand bootstrap replicates were sampled for the calculated statistical support. Thickened lines indicate branches with bootstrap values $\geq 70\%$. *C. neoformans* clinical strain H99 served as an outgroup. **b**, The complete sequence of the R265 common outbreak *MAT α* allele was compared to ~20 kb of sequence from coding and intergenic regions of the R272 outbreak isolate, revealing 0.37% sequence divergence.

of the major and minor genotypes revealed that 14 loci were identical and 16 were divergent, with 1–13 changes between the alleles in an average MLST length of 585 base pairs (bp). The probability that this distribution of mutations occurred by chance after genetic drift is highly unlikely (see Methods).

As the proportion of 14 identical:16 divergent alleles shared between the major and minor outbreak strains R265 and R272 is not significantly different from that expected by mendelian segregation of unlinked loci following meiosis ($P > 0.5$), the most parsimonious explanation is that the two Vancouver Island outbreak genotypes are either siblings from a genetic cross, or that one is the parent and the other a progeny. Given that the minor outbreak genotype also exists in multiple locations in Australia, but the major outbreak genotype has only been found in the Pacific Northwest, we favour the hypothesis that the minor genotype represents one of two parental strains that gave rise to the major outbreak isolate (Supplementary Information). Furthermore, the Vancouver Island major genotype bears multiple loci with alleles unique to this population, indicating that the second parent remains to be discovered.

Although the second parental isolate has not been accounted for, its relationship to the identified parent can be established through phylogenetic analysis. For all genes analysed, the phylogenies formed clades representing the four VG types (Fig. 1 and Supplementary Fig. 1). The two Vancouver Island genotypes always clustered within the VGII clade, including for the five MLST loci at which the two genotypes differ. Thus, the meiotic event from which the Vancouver Island major genotype descends occurred between two parental VGII strains. Furthermore, for multiple independent, non-recombining populations the genealogies of multiple genes for each strain should be equivalent. This holds true for all strains analysed. The absence of inter-VG hybrid genotypes indicates that meiotic events involving *C. gattii* tend to occur within, rather than between, different VG types.

Analysis of the mating-type locus provides evidence that the major outbreak isolate was not produced by a traditional α - α cross, and may instead be the product of an α - α sexual cycle¹¹. In a traditional sexual cycle, all α progeny inherit an identical α allele from the α parent (identity by descent)²¹. However, the *SXII* α alleles from the two Vancouver Island genotypes differ by a single nucleotide (Fig. 2a).

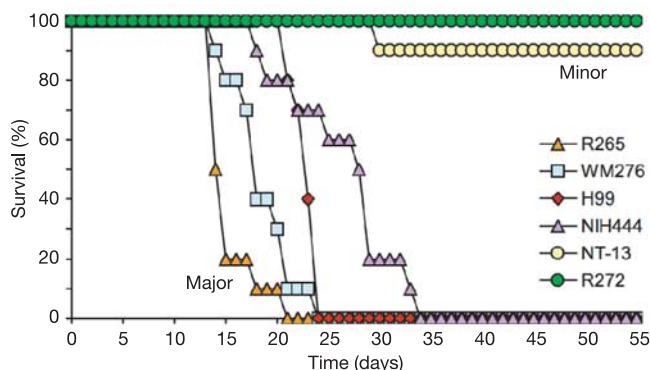


Figure 3 | The *C. gattii* major outbreak genotype is hypervirulent and the minor genotype is attenuated. The *C. gattii* R265 strain was significantly more virulent in a murine intranasal inhalation infection model compared with *C. neoformans* strain H99 ($P < 0.01$). Strain NIH444 showed reduced virulence compared with R265 ($P < 0.001$), but was not significantly different from strain H99 ($P \geq 0.05$). Strains R272 and NT-13 were significantly less virulent ($P < 0.001$). We note that NIH444 is less virulent than R265. This difference might be the result of laboratory passage or storage. Groups of ten A/Jcr mice were infected with 5×10^4 c.f.u. of the indicated *C. gattii* strain by intranasal inhalation. Statistical analysis was performed using the Mann–Whitney *U*-test.

The major genotype *SXII* α allele is present only in other strains of this genotype from Vancouver Island, as well as in NIH444, CBS7750 and ICB107 isolates (Supplementary Table 1), but not in any known Australian isolates¹⁹. To examine this apparent difference, the *MAT* α alleles of the entire collection were fingerprinted by amplifying variable intergenic regions. Although different strains yielded discrete *MAT* α intergenic region fingerprints, the two outbreak genotypes had an identical profile (Fig. 2a and Supplementary Table 3). Next, the complete locus from the major outbreak strain R265 was sequenced (100,201 bp) and compared with ~ 20 kb of *MAT* sequence for the minor outbreak strain R272 (Fig. 2b). The R265 and R272 sequences contain a widespread low level of polymorphism (0.37% divergence), establishing that the two bear different—but related— α alleles. These data support the hypothesis that the Vancouver Island major genotype was produced as a result of mating between two closely related α VGII parental strains, one of which we hypothesize is still present as the minor outbreak isolate and fertile Australian isolates.

One clue that the two Vancouver Island genotypes might differ in virulence was that although the two are closely related, they differ markedly in prevalence in clinical isolates. We addressed this by testing the virulence of representative isolates using a murine intranasal inhalation model. Two clinically derived isolates of each genotype (one from Vancouver Island, one from elsewhere) were tested, in addition to the divergent environmental *C. gattii* VGI strain WM276 and the *C. neoformans* clinical isolate H99. This analysis showed a striking difference in virulence for the two genotypes—the major outbreak genotype isolates (R265 and NIH444) were highly virulent, but the minor outbreak genotype and corresponding Australian isolate (R272 and NT-13, respectively) were avirulent or significantly attenuated (Fig. 3).

The Vancouver Island major genotype is identical to the type strain NIH444, a sample isolated in the 1970s (ref. 22). Why, then, has the outbreak occurred only recently? One explanation is that the parental and progeny strains were not introduced to Vancouver Island until a later date, and that the new progeny genotype is fitter in this environment. Consistent with this model, environmental sampling shows that the recombinant genotype is more prevalent on Vancouver Island. Sex within the same mating type might therefore confer evolutionary advantages when the opposite mating type is unavailable. In plants, transitions from outcrossing to self-pollination are common evolutionary events. In *Arabidopsis thaliana*, a selective sweep of a self-incompatibility pseudogene occurred concomitantly with the last ice age, and may have enabled global expansion by obviating the need for a sexual partner²³. Here, a similar transition might have contributed to expand the range of the fungal pathogen by enabling ready production of aerosolized spores. As the major genotype is ubiquitous in this environment, same-sex mating would occur most

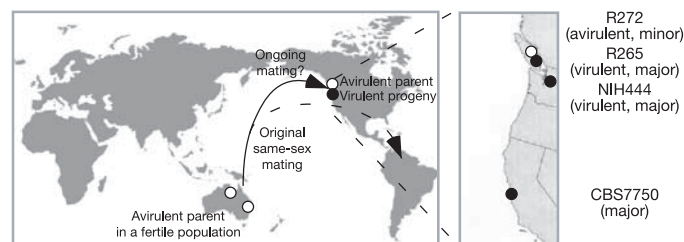


Figure 4 | The origin of an outbreak. Circles indicate the presence of the avirulent, parental genotype (open circles) and the virulent, recombinant genotype (filled circles). One α mating-type parent probably originated in Australia, and may have been introduced to the Pacific Northwest in the early twentieth century along with imported eucalyptus trees. An original α - α mating event yielded a virulent recombinant, and ongoing α - α mating may enable robust infectious spore production and outbreak expansion. Inset shows that the outbreak virulent genotype has also been isolated from Seattle and San Francisco.

commonly between isolates of identical genotype and therefore produce meiotically derived clones undetectable by molecular approaches. Monokaryotic fruiting has been reported in *C. gattii*²⁴, and studies reveal that fruiting represents a sexual cycle involving diploidization, meiosis and sporulation¹¹. Flow cytometric analysis of 95 outbreak strains revealed one diploid environmental isolate (RB59) homozygous for the major genotype (α/α), which may represent a same-sex mating intermediate. Therefore, ongoing same-sex mating could contribute to the production of infectious spores and further variants.

The emergence of a recombinant genotype representing the major form in the Vancouver Island outbreak was enabled by the presence of fertile parental isolates. Where did these strains originate? It seems likely that the parental, minor genotype derives from Australia, given that it shares an identical genotype with several fertile Australian isolates (Fig. 4). We identified the major genotype in only two locations other than Vancouver Island, one of these in San Francisco, associated with *Eucalyptus camaldulensis*, a tree species imported to the United States from Australia in the early twentieth century. We hypothesize that the minor genotype isolate mated with a second VGII isolate—in Australia, in transit or in the Pacific Northwest. If the meiotic event occurred in Australia, the major genotype may remain undiscovered on the sparsely populated continent. The major outbreak genotype might also be present in other locations where eucalypts have been exported, including South America, where we have identified similar genotypes (Supplementary Information).

As the second parental genotype has proven elusive, it is currently not possible to ascertain whether virulence of the recombinant outbreak isolate was enhanced by reassortment of two less pathogenic genomes, or by inheritance of specific parental virulence attributes. Altered pathogenicity in recombinant progeny has been illustrated in *T. gondii*, in which sexual recombination admixed two ancient lineages, giving rise to extant clones with enhanced oral transmission². Experimental *T. gondii* crosses show that sex can dramatically enhance virulence²⁵. Other human pathogens may harbour cryptic same-sex cycles that contribute to produce progeny with altered pathogenicity, host range and/or niche adaptation. This could operate in other largely clonal species, for which population studies have not detected genetic reassortment. This includes *Pneumocystis carinii*, in which synaptonemal complexes occur but only one mating type is known²⁶. In addition, same-sex mating could be common in parasites such as *Trypanosoma cruzi*, *Leishmania* and *Plasmodium falciparum*, in which parasexual and sexual reproduction transpire but mating-type determinants are unknown^{27,28}. Further studies will be required to establish the prevalence of same-sex *C. gattii* mating events in nature and whether this process is enabling robust spore production and expansion of the outbreak, in addition to an original α - α mating that yielded a new clone able to cause an infectious disease outbreak in humans.

METHODS

Strains. Strains are described in Supplementary Table 1. All strains were revived from 15% glycerol stocks stored at -80°C .

MLSTs and phylogenetic analysis. MLST loci were selected on different chromosomes or $>100,000$ nucleotides apart. MLST locus primers (Supplementary Table 4) were designed to amplify variable non-coding DNA regions to maximize strain discrimination. For all loci, both DNA strands were sequenced and edited manually using Sequencher 4.1 (Gene Codes Corporation) and assigned an allele number (Supplementary Tables 1 and 2). GenBank accession numbers for sequenced alleles are listed in Supplementary Table 5. Sequences in FASTA format were automatically aligned using CLUSTAL W²⁹ and manually edited in MacClade 4.06. Maximum parsimony analyses were performed using PAUP* 4.0b10. Maximum parsimony trees were rooted using *C. neoformans* strain H99. Statistical support for each data set was calculated with 1,000 parsimony bootstrap replicates.

To test whether the Vancouver Island major and minor genotypes could be related as a result of genetic drift, a Perl computer simulation was conducted. Genetic drift was simulated by concatenating sequenced loci, with mutations

randomly assigned to a base within the concatenated sequence with a uniform probability, and then classified by locus. In this way, the probability of a locus receiving a mutation was dependent on length and number of loci. Each simulation run assigned 59 mutations (the number of observed mutations) to 30 loci, and the observed number of loci with mutations was tallied. One hundred thousand replicates were run, and the number of configurations of 14 or more loci with no mutations established the probability that the data could be explained by genetic drift. Out of 100,000 replicates, the chance occurrence of 11 polymorphism-free loci was observed only once, but never seen in 12 or more loci, yielding a *P* value of <0.0001 and indicating that the observed relationship between R265 and R272 is unlikely to be explained by genetic drift. We repeated the simulation giving all loci equal lengths, and found an equivalent result.

MAT locus fingerprinting and sequencing. *MAT* α locus intergenic regions were amplified using polymerase chain reaction (PCR) primers based on the locus from isolate WM276 (ref. 16; Supplementary Table 3). Primer-binding sites were designed to protein-coding sequences; 16 regions were identified in and around *MAT* in WM276 that could be reproducibly amplified from each strain, and were separated on a 1% agarose TAE gel, visualized with ethidium bromide, and compared to size standards. Representative fragments were confirmed by sequencing as being *MAT*-derived. The R265 *MAT* α locus was isolated from a large-insert bacterial artificial chromosome (BAC) library from J. Kronstad. BAC clones were arrayed on nylon membranes, hybridized to *C. gattii* *MAT* probes, and a BAC containing the entire locus was sequenced as described¹⁶. The R272 *MAT* α sequence was determined from fragments amplified during fingerprinting. Supplementary Table 5 lists GenBank accession numbers for all new sequences.

Virulence studies. Virulence was assessed using female A/Jcr mice (NCI/Charles River Laboratories, 20–24 g). Strains were cultured in YPD broth for 18–20 h at 30°C , harvested, washed three times with sterile phosphate-buffered saline (PBS) and counted using a haemocytometer to determine cell concentrations. Inocula for all experiments were confirmed by plating on YPD and counting colony-forming units (c.f.u.). Ten A/Jcr mice per strain were infected by intranasal inhalation with 5×10^4 c.f.u. in $50 \mu\text{l}$ PBS (ref. 30). Animals that appeared moribund or in pain were killed. The Duke University Animal Use Committee approved this protocol. Mortality and time to mortality were evaluated for statistical significance both with the Mann–Whitney *U*-test and with Kaplan–Meier survival curves, and *P* values were obtained from a log-rank test. Both statistical analyses gave equivalent results.

Received 29 June; accepted 12 September 2005.

Published online 9 October 2005.

- Monot, M. et al. On the origin of leprosy. *Science* **308**, 1040–1042 (2005).
- Su, C. et al. Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* **299**, 414–416 (2003).
- Hoang, L. M., Maguire, J. A., Doyle, P., Fyfe, M. & Roscoe, D. L. *Cryptococcus neoformans* infections at Vancouver Hospital and Health Sciences Centre (1997–2002): epidemiology, microbiology and histopathology. *J. Med. Microbiol.* **53**, 935–940 (2004).
- Casadevall, A. & Perfect, J. R. *Cryptococcus neoformans* (ASM Press, Washington DC, 1998).
- Kwon-Chung, K. J., Boekhout, T., Fell, J. W. & Diaz, M. Proposal to conserve the name *Cryptococcus gattii* against *C. honduricus* and *C. bacillisporus* (Basidiomycota, Hymenomycetes, Tremellomycetidae). *Taxon* **51**, 804–806 (2002).
- Ruiz, A. & Bulmer, G. S. Particle size of airborne *Cryptococcus neoformans* in a tower. *Appl. Environ. Microbiol.* **41**, 1225–1229 (1981).
- Sukroongreung, S., Kitinyom, K., Nilakul, C. & Tantimavanich, S. Pathogenicity of basidiospores of *Filobasidiella neoformans* var. *neoformans*. *Med. Mycol.* **36**, 419–424 (1998).
- Hull, C. M. & Heitman, J. Genetics of *Cryptococcus neoformans*. *Annu. Rev. Genet.* **36**, 557–615 (2002).
- Fraser, J. A., Subaran, R. L., Nichols, C. B. & Heitman, J. Recapitulation of the sexual cycle of the primary fungal pathogen *Cryptococcus neoformans* var. *gattii*: implications for an outbreak on Vancouver Island, Canada. *Eukaryot. Cell* **2**, 1036–1045 (2003).
- Kidd, S. E. et al. A rare genotype of *Cryptococcus gattii* caused the cryptococcosis outbreak on Vancouver Island (British Columbia, Canada). *Proc. Natl Acad. Sci. USA* **101**, 17258–17263 (2004).
- Lin, X., Hull, C. M. & Heitman, J. Sexual reproduction between partners of the same mating type in *Cryptococcus neoformans*. *Nature* **434**, 1017–1021 (2005).
- Boekhout, T. et al. Hybrid genotypes in the pathogenic yeast *Cryptococcus neoformans*. *Microbiol.* **147**, 891–907 (2001).
- Litvitseva, A. P., Thakur, R., Vilgalys, R. & Mitchell, T. G. Multilocus sequence typing reveals three genetically distinct subpopulations of *Cryptococcus neoformans* var. *grubii* (serotype A), including a unique population in Botswana. *Genetics* (in the press).

14. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145 (1998).
15. Hull, C. M., Boily, M. J. & Heitman, J. Sex-specific homeodomain proteins Sxi1 α and Sxi2 α coordinately regulate sexual development in *Cryptococcus neoformans*. *Eukaryot. Cell* **4**, 526–535 (2005).
16. Fraser, J. A. *et al.* Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biol.* **2**, 2243–2255 (2004).
17. Diaz, M. R., Boekhout, T., Theelen, B. & Fell, J. W. Molecular sequence analyses of the intergenic spacer (IGS) associated with rDNA of the two varieties of the pathogenic yeast, *Cryptococcus neoformans*. *Syst. Appl. Microbiol.* **23**, 535–545 (2000).
18. Campbell, L. T. *et al.* Clonality and recombination in genetically differentiated subgroups of *Cryptococcus gattii*. *Eukaryot. Cell* **4**, 1403–1409 (2005).
19. Campbell, L. T. *et al.* Clinical and environmental isolates of *Cryptococcus gattii* from Australia that retain sexual fecundity. *Eukaryot. Cell* **4**, 1410–1419 (2005).
20. Halliday, C. L. & Carter, D. A. Clonal reproduction and limited dispersal in an environmental population of *Cryptococcus neoformans* var. *gattii* isolates from Australia. *J. Clin. Microbiol.* **41**, 703–711 (2003).
21. Loftus, B. J. *et al.* The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324 (2005).
22. Kwon-Chung, K. J. A new species of *Filobasidiella*, the sexual state of *Cryptococcus neoformans* B and C serotypes. *Mycologia* **68**, 943–946 (1976).
23. Shimizu, K. K. *et al.* Darwinian selection on a selfing locus. *Science* **306**, 2081–2084 (2004).
24. Wickes, B. L., Mayorga, M. E., Edman, U. & Edman, J. C. Dimorphism and haploid fruiting in *Cryptococcus neoformans*: association with the α -mating type. *Proc. Natl Acad. Sci. USA* **93**, 7327–7331 (1996).
25. Grigg, M. E., Bonnefoy, S., Hehl, A. B., Suzuki, Y. & Boothroyd, J. C. Success and virulence in *Toxoplasma* as the result of sexual recombination between two distinct ancestries. *Science* **294**, 161–165 (2001).
26. Thomas, C. F. Jr & Limper, A. H. *Pneumocystis pneumonia*. *N. Engl. J. Med.* **350**, 2487–2498 (2004).
27. Gaunt, M. W. *et al.* Mechanism of genetic exchange in American trypanosomes. *Nature* **421**, 936–939 (2003).
28. Alano, P. & Carter, R. Sexual differentiation in malaria parasites. *Annu. Rev. Microbiol.* **44**, 429–449 (1990).
29. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
30. Cox, G. M. *et al.* Superoxide dismutase influences the virulence of *Cryptococcus neoformans* by affecting growth within macrophages. *Infect. Immun.* **71**, 173–180 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Kronstad for the R265 BAC library and strains; A. Litvintseva for MLST information; F. Dromer, D. Carter and C. Paula for strains; A. Idnurm, C. Fraser, X. Lin, K. Nielsen, J. Blankenship, A. Litvintseva, D. Lew and J. Anderson for reading the manuscript; and the Broad Fungal Genome Initiative. This work was supported by an NIAID R01 grant to J.H.

Author Contributions J.A.F. and J.H. designed the experiments and wrote the manuscript. J.A.F. conducted or participated in all of the experiments, and S.S.G. and S.G.G-B. assisted with animal experiments. J.R.W. and J.R.P. supervised animal experiments and assisted with data analysis. E.C.W. contributed technical support. S.D. assisted with phylogenetic analysis. A.A. and F.S.D. contributed sequence analyses for MLST and MAT, and J.E.S. conducted computer modelling for genetic drift.

Author Information DNA sequences identified in this study have been deposited in GenBank; a full list of accession numbers is available in Supplementary Table 5. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.H. (heitm001@duke.edu).

Mapping determinants of human gene expression by regional and genome-wide association

Vivian G. Cheung^{1,2,3}, Richard S. Spielman², Kathryn G. Ewens², Teresa M. Weber^{2,3}, Michael Morley³ & Joshua T. Burdick³

To study the genetic basis of natural variation in gene expression, we previously carried out genome-wide linkage analysis and mapped the determinants of ~1,000 expression phenotypes¹. In the present study, we carried out association analysis with dense sets of single-nucleotide polymorphism (SNP) markers from the International HapMap Project². For 374 phenotypes, the association study was performed with markers only from regions with strong linkage evidence; these regions all mapped close to the expressed gene. For a subset of 27 phenotypes, analysis of genome-wide association was performed with >770,000 markers. The association analysis with markers under the linkage peaks confirmed the linkage results and narrowed the candidate regulatory regions for many phenotypes with strong linkage evidence. The genome-wide association analysis yielded highly significant results that point to the same locations as the genome scans for about 50% of the phenotypes. For one candidate determinant, we carried out functional analyses and confirmed the variation in *cis*-acting regulatory activity. Our findings suggest that association studies with dense SNP maps will identify susceptibility loci or other determinants for some complex traits or diseases.

The expression level of genes, 'the expression phenotype'³, is highly variable and heritable in humans^{4,5} and other organisms⁶. We previously¹ performed genome-wide linkage analysis for 3,554 expression phenotypes in 14 pedigrees from the Utah component of the Centre d'Etude du Polymorphisme Humain (CEPH) to map the genetic determinants of variation in human gene expression. For ~1,000 expression phenotypes, significant linkage evidence was obtained, suggesting the existence of determinants that act in *cis* or *trans* to the expressed gene. In the present study, we used genotype data generated by the International HapMap Project² to test for allelic association in two sets of analyses. First, for a set of 374 phenotypes with evidence of *cis*-linked determinants, we performed association analysis with dense sets of SNPs near linkage peaks. Second, we restricted attention to the 27 phenotypes with the strongest linkage evidence for *cis*-acting determinants, and tested >770,000 SNPs in the genome for association. Because all the phenotypes have been mapped previously by genetic linkage, the analyses allow us to compare directly the results from whole-genome linkage and association studies.

The 374 phenotypes have at least one marker with linkage evidence ($t > 2$) for *cis* regulators¹; this corresponds to a point-wise $P < 0.02$ with the sample of 14 CEPH sibships. For the present analysis, we obtained SNP genotype data on 57 unrelated CEPH individuals from the International HapMap Project² and generated expression phenotypes using the Affymetrix Human Genome Focus arrays.

Evidence for linkage requires co-segregation between the phenotype and a marker site, but does not depend on the particular allele

present at the marker. In contrast, allelic association with a linked marker requires correlation with a particular SNP allele; that is, linkage disequilibrium. Even if there are several different alleles at the determinant ('allelic heterogeneity'), linkage can be detected. But if there is allelic heterogeneity, it is less likely that there will be detectable association. Therefore, it was not obvious that evidence for linkage would predict evidence for association. So, for a set of phenotypes with *cis* linkage, we performed association analysis with SNPs within the target genes and within 50 kilobases (kb) of the 5' and 3' ends, and compared results with those from the previous linkage scans¹. The evidence for association was assessed by linear regression. Among the 374 phenotypes, there are 65 (17%) with at least one marker that shows evidence of association at the nominal $P < 0.001$ level. For some of the phenotypes, the association with a nearby marker is extremely strong; among the 65 phenotypes, there are 12 with evidence of *cis* association at $P < 10^{-10}$. At the less stringent threshold of $P < 0.01$ for association, there are 133 (36%) phenotypes. We also determined the proportion of phenotypes with these two nominal levels of evidence for *cis* association for various strengths of initial linkage findings (Supplementary Table 1). We found that the strength of linkage evidence did tend to predict association results. For example, among the 27 phenotypes with highly significant *cis* linkage ($t > 5$, $P < 3.7 \times 10^{-5}$), 70% have evidence of *cis* association at $P < 0.001$, compared to only 9% of the phenotypes with modest evidence of *cis* linkage ($2 < t < 3$, $P < 0.02$).

Although there are many examples of regulatory sites located in 5' or 3' flanking regions of genes, little is known about the relative frequencies. Although the marker most strongly associated with gene expression level is not necessarily the functional variant, we expect that in most cases that marker will be very close to the functional variant. With this assumption, we determined the location of the markers within 50 kb of the target genes that showed the strongest association, to establish whether they occur preferentially in the 5' or 3' regions. Among the 133 phenotypes with *cis* association at $P < 0.01$, the regulatory sites are found in approximately the same proportions in the 5' (27%) and 3' (34%) ends, and within the target genes (25%). For 14% of the phenotypes, linkage disequilibrium among SNPs spanning the regions examined was so strong that we were not able to narrow the regions of *cis* association. Thus, overall we found that *cis* regulators are not preferentially enriched in the 5' or 3' regions around the target genes. However, for most of the phenotypes, the analysis of regional association data narrowed the search for the regulatory determinants to one particular region near the target gene.

The analyses described so far were restricted to the SNPs known to be located in regions near the target genes. If we did not know in advance where to look for determinants, how successfully would we

¹Department of Pediatrics and ²Department of Genetics, University of Pennsylvania, ³The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA.

Table 1 | Genome-wide association results for 27 phenotypes

Phenotype	Location of target gene	Linkage results		GWA results (for peak marker)	
		Peak marker	<i>P</i> -value (all <i>cis</i>)	Marker	Location* Nominal <i>P</i> -value†
LRAP (LOC64167)	5q15		1×10^{-7}	rs2762	58,030 1.98×10^{-19}
AA827892	20q11.23		3×10^{-8}	rs788350	-666 3.67×10^{-15}
PSPHL	7p11.2		3×10^{-11}	rs6593279	-36,903 9.59×10^{-15}
CPNE1	20q11.22		1×10^{-7}	rs6060535	17,327‡ 8.35×10^{-13}
CSTB	21q22.3		2×10^{-9}	rs880987	-28,195 2.48×10^{-12}
RPS26	12q13.2		2×10^{-9}	rs2271194	-41,768 7.94×10^{-12}
GSTM2	1p13.3		3×10^{-8}	rs535088	12,699 2.00×10^{-11}
HLA-DRB2	6p21.32		$<10^{-11}$	rs6928482	8,345 6.51×10^{-11}
IRF5	7q32.1		2×10^{-8}	rs2280714	16,731 6.78×10^{-11}
HSD17B12	11p11.2		2×10^{-11}	rs4755741	100,949‡ 7.38×10^{-11}
GSTM1	1p13.3		1×10^{-7}	rs535088	-7,052 8.33×10^{-10}
PPAT	4q12		2×10^{-7}	rs227940	<i>Trans</i> (Chr 7) 5.29×10^{-9}
PPAT	4q12		2×10^{-7}	rs2139512	25,227‡ 2.87×10^{-8}
DDX17	22q13.1		6×10^{-10}	rs10490570	<i>Trans</i> (Chr 2) 7.13×10^{-9}
CTSH	15q25.1		7×10^{-9}	rs1369324	-2,298 2.17×10^{-8}
POMZP3	7q11.23		9×10^{-10}	rs1754162	-6,215 7.23×10^{-8}
CGI-96	22q13.2		3×10^{-9}	rs9600337	<i>Trans</i> (Chr 13) 2.43×10^{-7}
CHI3L2	1p13.3		3×10^{-11}	rs755467	-91 2.57×10^{-7}
VAMP8	2p11.2		9×10^{-8}	rs10509846	<i>Trans</i> (Chr 10) 5.31×10^{-7}
EIF3S8	16p11.2		4×10^{-8}	rs8092794	<i>Trans</i> (Chr 18) 7.20×10^{-7}
TM7SF3	12p11.23		$<10^{-11}$	rs11822822	<i>Trans</i> (Chr 11) 7.32×10^{-7}
IL16	15q25.1		3×10^{-10}	rs6957902	<i>Trans</i> (Chr 7) 9.63×10^{-7}
TCEA1	8q11.23		6×10^{-8}	rs6562160	<i>Trans</i> (Chr 13) 1.08×10^{-6}
S100A13	1q21.3		3×10^{-8}	rs3757791	<i>Trans</i> (Chr 7) 1.40×10^{-6}
ICAP-1A	2p25.1		$<10^{-11}$	rs10807387	<i>Trans</i> (Chr 6) 2.27×10^{-6}
SMARCB1	22q11.23		4×10^{-7}	rs7802273	<i>Trans</i> (Chr 7) 2.46×10^{-6}
CTBP1	4p16.3		2×10^{-9}	rs1060043	<i>Trans</i> (Chr 19) 5.26×10^{-6}
ZNF85	19p12		9×10^{-9}	rs2168903	<i>Trans</i> (Chr 12) 6.51×10^{-6}

* Relative to transcriptional start site of target gene. When the most significant marker is located on a chromosome different from the target gene, it is listed as 'Trans' and the chromosome is shown.

† Corrected *P*-value of 0.05 corresponds to a nominal *P*-value of 6.7×10^{-8} .

‡ Marker is within genomic extent of target gene.

find them? To answer this question, we took advantage of the hundreds of thousands of markers across the genome, genotyped on the same 57 unrelated CEPH individuals as above. Instead of focusing on *cis* regulatory regions, we performed genome-wide association analysis (GWA) to map determinants. We limited our analysis to the 27 phenotypes with the strongest evidence of *cis* regulation from our whole-genome linkage analysis¹ so that the results could be compared with the linkage results in which we have highest confidence. We tested 770,394 SNP markers for association with each expression phenotype and performed a regression analysis of expression level on marker genotype. We used the Šidák procedure, which is conservative⁷, to correct for multiple testing.

The evidence for association was significant at the genome-wide level (nominal $P < 6.7 \times 10^{-8}$ or corrected $P < 0.05$) for 14 of the 27 phenotypes (Table 1). Figure 1 shows several examples. The GWA identified only *cis* regulation for 12 of the phenotypes, *cis* and *trans* regulation for another phenotype (phosphoribosyl pyrophosphate aminotransferase (*PPAT*)), and only *trans* regulation for one phenotype (DEAD box polypeptide 17 (*DDX17*)).

We compared the findings for the 27 phenotypes in Table 1 to those from our previous whole-genome linkage scans¹. To simplify this analysis, we focused on the SNP with the most significant finding in the GWA for each phenotype in Table 1. For this SNP, we classified the results according to whether the association was or was not supported by the genome-wide linkage analysis (that is, whether the SNP that showed significant association fell within the region of significant linkage). For 15 of the 27 phenotypes, genome-wide linkage analyses and GWA pointed to the same *cis* regulatory regions. In 13 of these, the corrected *P*-value for the *cis* marker was <0.05 ($P_{\text{uncorr.}} < 6.7 \times 10^{-8}$) in the GWA, providing highly significant evidence for *cis* regulatory elements. In one of the 13 phenotypes, *PPAT*, the GWA result points to both *cis* and *trans* regulators; however, linkage scan results support only the *cis* regulation. For the other two phenotypes (*POMZP3* and *CHI3L2*) the results were

nominally significant ($P = 7 \times 10^{-8}$ and $P = 3 \times 10^{-7}$, respectively). For both phenotypes, the peak SNPs are located close to the target genes (6 kb 5' for *POMZP3* and 91 base pairs (bp) 5' for *CHI3L2*). Because of the location, we consider these cases 'true' positives, but recognize that the statistical evidence alone is not compelling. We show evidence below that one of these peak SNPs is a regulatory variant of *CHI3L2*.

For the remaining 12 of the 27 phenotypes, no strong evidence for

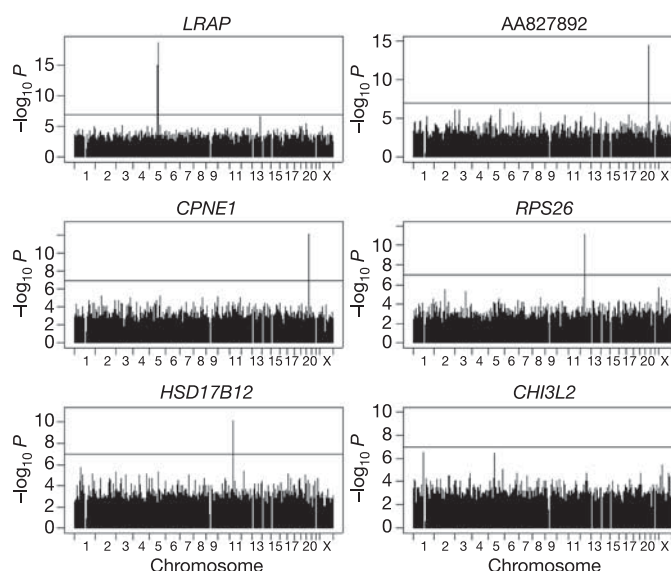


Figure 1 | Results of genome-wide association analysis for six representative phenotypes with *cis* regulators. The horizontal line in each panel corresponds to $P = 0.05$ after Šidák correction.

cis determinants was found by GWA. In 11 of the 12 phenotypes, no significant association ($P_{\text{corr.}} < 0.05$ or $P_{\text{uncorr.}} < 6.7 \times 10^{-8}$) was detected anywhere in the genome, even though highly significant evidence for *cis* regulation was detected in the linkage scans. For *DDX17*, as indicated above, GWA identified significant *trans* association ($P_{\text{corr.}} < 0.05$) but evidence for a *trans*-acting regulator was not found in the linkage scan. For all 12 phenotypes, the most significant markers were located on chromosomes different from that of the expressed gene. (There are 13 '*trans*' markers listed in Table 1; one is the non-*cis* marker for *PPAT*, see above.) The locations and *P*-values from the association analyses for these markers are shown in Table 1. In some cases, the failure to find significant evidence for *cis* association, despite highly significant linkage results for *cis* regulation, may be due to linkage findings that were false-positive results. However, even if the linkage findings reflect real effects, allelic heterogeneity might make the determinant undetectable by association. Furthermore, our sample size of 57 might be too small to yield statistically significant associations, especially if the marker closest to a determinant is not in strong linkage disequilibrium with it.

To assess the issue of sample size for this and future studies, we determined the power to detect association—with the same correction we used for multiple testing—under various assumptions about sample size. We also considered variation in effect size; that is, the proportion (R^2) of phenotype variation accounted for by the SNP. As a point of reference, the observed values of R^2 for the 13 phenotypes with significant *cis* association range from 0.44 to 0.78. (In classical complex traits and diseases, R^2 for individual determinants is expected to be much smaller.) To achieve a probability of 0.8 of detecting a determinant for which R^2 is 0.1, we found that samples of ~500 would be needed (see Supplementary Information). Even these results must be viewed as optimistic; the calculations apply to an idealized determinant that is identical to the marker in both location and frequency, yielding maximum linkage disequilibrium. Furthermore, it is difficult to extrapolate the results to complex traits in general. For many quantitative traits of interest, the fraction of phenotypic variance attributable to any one determinant is likely to be far smaller than the values we estimated for the 13 strongly associated *cis*-linked expression phenotypes. Thus, extrapolations from the small samples used here to genome-wide analysis with other

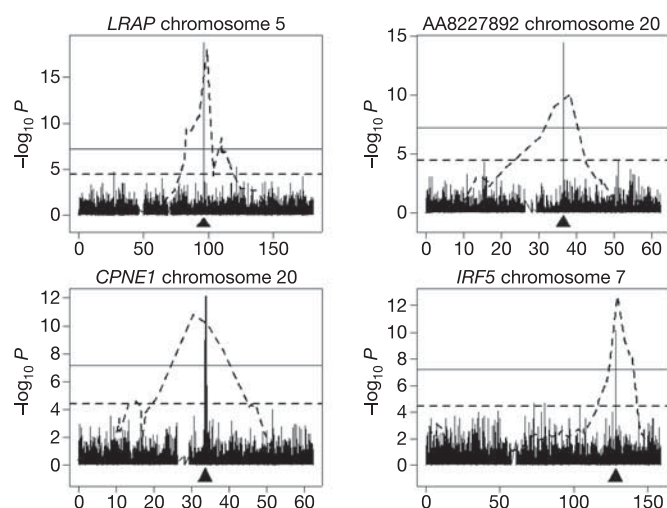


Figure 2 | Results of genome-wide linkage analysis (dotted line) superimposed on those from genome-wide association (bars) for the chromosome where the expressed gene is located. The location of the expressed gene is indicated by an arrow. The dotted horizontal line is for data from linkage scans and corresponds to $t = 4$, $P = 3.7 \times 10^{-5}$. The solid horizontal line is for data from GWA and corresponds to $P = 0.05$ after Šidák correction. The *x* axis indicates chromosome location in megabases.

complex traits should be done cautiously—much larger samples will usually be required.

We expected that the much greater resolution afforded by linkage disequilibrium would result in candidate regions defined by association that are much smaller than those defined by linkage. For all 15 phenotypes where the linkage and GWA results are concordant, this expectation was confirmed. Figure 2 shows several examples. For example, in the GWA for *CPNE1*, the 30 markers with significant evidence of association ($P_{\text{corr.}} < 0.05$) span a region of only 402 kb. For *IRF5*, the ten significant markers span 90 kb of chromosome 7. For four phenotypes (*RPS26*, *GSTM1*, *GSTM2* and *DDX17*), there is only one significant marker.

Ultimately, the findings from linkage and association need to be confirmed by functional analysis. Results from our genome-wide linkage and association analyses suggest that expression level of chitinase 3-like 2 (*CHI3L2*) is regulated by a *cis*-acting transcriptional regulator (Table 1). A marker (rs755467) in the promoter region of *CHI3L2* shows nominally significant association results ($P < 3 \times 10^{-7}$). Although it is not significant after correction for 770,394 markers, we considered it likely that expression level of *CHI3L2* is indeed regulated by a polymorphism in its promoter region, because both linkage and association results point to the same candidate regulatory region. We followed up this finding with functional tests. To evaluate the relative promoter activity of the haplotypes bearing the different alleles of the SNP marker rs755467, we performed luciferase reporter assays. Stronger luciferase expression (>3-fold) was observed for vectors containing the upstream fragment than for the empty vector, confirming promoter activity in the DNA fragment containing marker rs755467. We then compared the promoter activity of the fragment bearing the rs755467 T-allele with that bearing the G-allele. The promoter activity of the fragment bearing the rs755467 T-allele was 2 ± 0.6 -fold higher ($P < 0.05$) than that bearing the G-allele.

Next, we performed haplotype-specific chromatin immunoprecipitation (haploChIP) to determine whether this observation is due to differential binding of RNA polymerase II to the different haplotypes^{8,9}. The assay was performed using lymphoblastoid cells from three CEPH individuals heterozygous at rs755467. We used antibody against RNA polymerase II (phosphorylated serine 5), which is known to accumulate in promoter regions of genes¹⁰. We quantified the relative abundance of the RNA polymerase II binding to the haplotypes bearing the T- or G-alleles in each of the

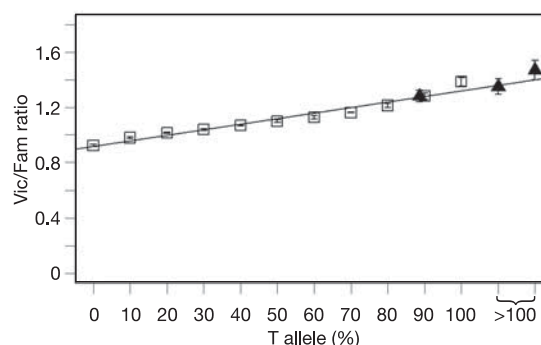


Figure 3 | HaploChIP assay with RNA polymerase II antibody to estimate allele-specific binding. Open squares show fluorescence intensities (Vic/Fam ratios) of genomic DNA samples with experimentally varied proportions of allele T at SNP rs755467. The linear regression line is shown. Solid triangles show mean fluorescence intensities of immunoprecipitated products from three individuals heterozygous at rs755467. In all three cases, the immunoprecipitated products contained more T than G in the DNA. Estimates greater than 100% for the T-allele imply that no G-allele-bearing DNA fragments were detected in the immunoprecipitated products. Standard deviation of triplicate measurements is shown.

heterozygous individuals by quantitative polymerase chain reaction (PCR)⁹. In all three cases, we found that in immunoprecipitated DNA, the amount of DNA bearing the T-allele of rs755467 is much higher than that bearing the G-allele ($\geq 90\%$ T and 10% G) (Fig. 3). This result suggests that the higher expression of *CHI3L2* in individuals who are TT homozygotes for rs755467 is due to stronger binding of RNA polymerase II to DNA containing the T-allele than to that containing the G-allele. Thus, results from both functional assays support data from the association analysis and suggest that the polymorphism rs755467 or a marker in strong linkage disequilibrium with it is a determinant that affects the expression level of *CHI3L2*.

The results of this study have implications for studies of many phenotypes besides gene expression. It has become a truism that finding genes that contribute to complex traits and diseases is much more difficult than finding genes for mendelian traits, and the underlying problems have been discussed in many reviews^{11–14}. As genomic resources have become available, increasing attention has been given to the role of association analysis, first described systematically in ref. 15. Here we have carried out association analysis with dense SNP maps in two settings that parallel approaches used to map genes for human diseases and traits. First, we performed regional association analysis with SNP markers in and around target genes to identify *cis*-acting regulatory variants. This approach parallels candidate gene studies. Second, we analysed >770,000 SNP markers to identify determinants of variation in gene expression by GWA analysis. This parallels studies where no candidate regions are specified in advance, and association analysis is used to find susceptibility genes. The GWA analysis, carried out with only 57 unrelated individuals, yielded highly significant results for 14 of the 27 phenotypes. The evidence for association was so strong in these cases that the findings remained significant ($P_{\text{corr.}} < 0.05$) even after correction for the large number of tests genome-wide.

The use of GWA to find genes is controversial^{12,16,17}. The optimism inspired by the availability of very dense SNP maps has been tempered by the resulting problems of multiple testing. There have been a few genome-wide association studies^{18–20}; however, the marker density in those studies was much lower than that made possible here by the International HapMap Project. Despite the small sample size, we obtained significant findings for approximately 50% of the 27 phenotypes analysed by genome-wide association analysis. The results suggest that as denser marker maps become available and high-throughput genotyping technologies are developed, large-scale association studies will become a practical means to identify genes for complex traits and diseases.

METHODS

CEPH samples and expression phenotyping. The data were from 57 CEPH individuals. These correspond to all the unrelated individuals in the CEPH HapMap collection except for three individuals (GM12264, GM11840, GM12056), because their cell lines were not available at the time of this study. For expression analysis, RNA was extracted from lymphoblastoid cells from each individual and hybridized onto Affymetrix Human Genome Focus arrays following the manufacturer's protocol. Expression intensity was scaled to 500 using Affymetrix Microarray Suite 5.0 program and $\log_2$ transformed.

Genotypes. SNP genotypes were downloaded from the International HapMap database (<http://www.hapmap.org>) (release 14). For data analysis, 770,394 markers were used; this excludes markers that are monomorphic or genotypes missing in all the 57 subjects. Locations of the SNP markers are based on those of the human reference sequence (<http://genome.ucsc.edu>) of May 2004 (hg 17, build 35).

Association analysis. For the association analysis, $\log_2$ -transformed expression level of the 57 individuals as a dependent variable was regressed on SNP genotype (coded 0, 1, 2). R^2 was estimated for each phenotype/genotype combination as the ratio of the regression sum of squares to total sum of squares.

For the association analysis with SNPs near and within the target genes, we have not used multiple-testing correction for the P -values (0.001 and 0.01) used as thresholds. Our goal is not to show that these are significant themselves, but to

compare the evidence from association and linkage analyses. The number of markers examined per phenotype varied depending on the genomic extent of the target gene. The median size of the target genes is 24 kb (mean = 45 kb). The average marker density in the version (release 14) of the HapMap data that we obtained is about 1 per 5 kb. Because the target genes vary in size, on average more markers were tested for the larger target genes. However, the genes with *cis* association were not significantly larger ($P > 0.5$) than those without *cis* association.

For the GWA, correction for multiple testing was performed by the Šidák procedure for the 770,394 tests⁷. The corrected P -value of 0.05 corresponds to a nominal P -value of 6.7×10^{-8} .

Power. The power of the association analysis for one SNP was estimated in two ways. First we carried out analyses on simulated data. We fixed total phenotypic variance ($= 1$, for convenience), and considered a range of values for R^2 . For each R^2 , we calculated corresponding differences between genotype means for the three SNP genotypes (assuming no dominance). Individual simulated phenotype values were obtained from each of three normal distributions, in numbers drawn from Hardy–Weinberg proportions, to reach the desired total sample size (60, 100, etc.). The process was repeated 500 times, and standard linear regression analysis was done. The proportion of outcomes with P -value $< 0.05/770,000$ was taken as the power. Second, this procedure was checked by comparing results with those from a published equivalent²¹.

Luciferase reporter assay. PCR was performed to generate fragments upstream of *CHI3L2* (–1,115 bp to +391 bp relative to transcription start site with forward primer 5'-ACGGAACGCGTTGCAAGTCCATTCTCACAGG-3' and reverse primer 5'-ACGGACTCGAGGACTCGCAATACAACGCTCA-3'). Genomic DNA from individuals homozygous at SNP marker rs755467 was used: GM12003 (GG), GM12236 (GG), GM12814 (TT) and GM10846 (TT). The amplicons were inserted into firefly luciferase reporter plasmids, pGL3-enhancer vectors (Promega). The clones were confirmed by sequencing. Two individuals of each genotype were examined as technical replicates. The pGL-enhancer reporter plasmids were co-transfected with pRL-TK *Renilla* control plasmids into lymphoblastoid cells (3 million cells per reaction) using Solution V (Amaya) by electroporation (Program T15) following the manufacturer's recommendation (Amaya). The cells were incubated for 24 h before lysis. Luciferase activity was read using a luminometer (Veritas) and normalized to the intensity of pRL-TK and reported as relative luciferase activity. All assays were performed with six replicates.

Quantitative HaploChIP. Chromatin immunoprecipitation was carried out in a similar manner as previously described (ref. 22; see also <http://genomics.ucdavis.edu/farnham/protocols/chips.html>). Cell lysates were sonicated (ten times for 10 s) using a Branson Ultrasonic Sonicator, resulting in DNA fragments with an average length of <500 bp. Lymphoblastoid cells (20×10^6 cells per reaction) from individuals (GM07000, GM07029 and GM10847) heterozygous at SNP marker rs755467, and antibody against 5Ser-P RNA polymerase II (clone H14, Covance), were used.

Allele-specific quantification of the immunoprecipitated product was carried out as previously described⁹ with Taqman primers for rs755467 (Applied Biosystems) using an ABI PRISM 7000. A standard curve for the quantitative PCR was produced with genomic DNA mixtures with varying amounts of DNA from two individuals that were homozygous at marker rs755467 (GM12814, TT; GM11829, GG). Using the standard curve, genomic DNA from two individuals (GM07029, GM10847) heterozygous at rs755467 was found to contain ~50% of T- and 50% of G-alleles as expected (reporter dye Vic (allele T) to reporter dye Fam (allele G) (Vic/Fam) ratio = 1.07). The standard curve was then used to quantify the ratio of T:G alleles in the immunoprecipitated DNA.

Received 30 June; accepted 19 September 2005.

1. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
2. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
3. Cheung, V. G. & Spielman, R. S. The genetics of variation in gene expression. *Nature Genet.* **32**, 522–525 (2002).
4. Cheung, V. G. et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genet.* **33**, 422–425 (2003).
5. Monks, S. A. et al. Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
6. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
7. Westfall, P. H. & Young, S. S. *Resampling-Based Multiple Testing* (John Wiley & Sons, New York, 1992).
8. Knight, J. C., Keating, B. J., Rockett, K. A. & Kwiatkowski, D. P. *In vivo* characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nature Genet.* **33**, 469–475 (2003).

9. Liu, X. *et al.* Expression-based discovery of variation in the human glutathione S-transferase M3 promoter and functional analysis in a glioma cell line using allele-specific chromatin immunoprecipitation. *Cancer Res.* **65**, 99–104 (2005).
10. Komarnitsky, P., Cho, E. J. & Buratowski, S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.* **14**, 2452–2460 (2000).
11. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
12. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
13. Glazier, A. M., Nadeau, J. H. & Aitman, T. J. Finding genes that underlie complex traits. *Science* **298**, 2345–2349 (2002).
14. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33**, 228–237 (2003).
15. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
16. Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452 (2004).
17. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
18. Carrasquillo, M. M. *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genet.* **32**, 237–244 (2002).
19. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
20. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
21. Ball, R. D. Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics* **170**, 859–873 (2005).
22. Oberley, M. J., Tsao, J., Yau, P. & Farnham, P. J. High-throughput screening of chromatin immunoprecipitates using CpG-island microarrays. *Methods Enzymol.* **376**, 315–334 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. J. Ewens, H. H. Kazazian and D. Bell for discussions and advice. This work is supported by grants from the National Institutes of Health (to R.S.S. and V.G.C.) and the W.W. Smith Endowed Chair (to V.G.C.).

Author Information The GEO accession number for the microarray data is GSE2552. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.G.C. (vcchung@mail.med.upenn.edu) or R.S.S. (spielman@pobox.upenn.edu).

LETTERS

Wnt signalling regulates adult hippocampal neurogenesis

Dieter-Chichung Lie^{1,2*}, Sophia A. Colamarino^{1*}, Hong-Jun Song^{1,3}, Laurent Désiré¹, Helena Mira¹, Antonella Consiglio¹, Edward S. Lein¹, Sebastian Jessberger¹, Heather Lansford¹, Alejandro R. Dearie¹ & Fred H. Gage¹

The generation of new neurons from neural stem cells is restricted to two regions of the adult mammalian central nervous system: the subventricular zone of the lateral ventricle, and the subgranular zone of the hippocampal dentate gyrus¹. In both regions, signals provided by the microenvironment regulate the maintenance, proliferation and neuronal fate commitment of the local stem cell population¹. The identity of these signals is largely unknown. Here we show that adult hippocampal stem/progenitor cells (AHPs) express receptors and signalling components for Wnt proteins, which are key regulators of neural stem cell behaviour in embryonic development². We also show that the Wnt/ β -catenin pathway is active and that Wnt3 is expressed in the hippocampal neurogenic niche. Overexpression of Wnt3 is sufficient to increase neurogenesis from AHPs *in vitro* and *in vivo*. By contrast, blockade of Wnt signalling reduces neurogenesis from AHPs *in vitro* and abolishes neurogenesis almost completely *in vivo*. Our data show that Wnt signalling is a principal regulator of adult hippocampal neurogenesis and provide evidence that Wnt proteins have a role in adult hippocampal function.

Using *in situ* hybridization, we found that Wnt3 is expressed in close proximity to the subgranular zone (SGZ), the region in which AHPs proliferate and differentiate into dentate granule neurons¹ (Fig. 1a). Analysis of adult Wnt/ β -catenin reporter mice (BATGAL)³ showed that this pathway is active in the SGZ and the dentate granule cell layer (Fig. 1b). We have previously shown that factors derived from the local astrocyte population participate in the regulation of proliferation and neuronal differentiation in hippocampal neurogenesis⁴. Notably, we found that several Wnt family members, including Wnt3, are expressed in adult hippocampal astrocytes (Fig. 1c and data not shown) and that AHPs express receptors for Wnts and crucial Wnt/ β -catenin signalling pathway components⁵ (Fig. 1d and data not shown). On the basis of these expression patterns, we considered that astrocyte-derived Wnts and Wnt signalling might be involved in the regulation of hippocampal neurogenesis.

We first tested this hypothesis in a co-culture system of AHPs and adult hippocampal astrocytes that models the interaction of these cell populations in the hippocampal neurogenic niche⁴. AHPs expressing green fluorescent protein (GFP) were co-cultured for 4 d with hippocampal astrocytes in the presence or absence of a Wnt inhibitor, secreted Frizzled-related protein 2 and 3 (sFRP2/3)⁶. Neuronal differentiation was evaluated every 24 h by quantifying the percentage of AHPs expressing the immature neuronal marker doublecortin⁷ (DCX; Fig. 1e). We also determined the percentage of AHPs that expressed the mature neuronal marker MAP2ab⁴ after

4 d (Fig. 1f). At all time points, the percentage of AHPs that differentiated into neurons was significantly decreased in the presence of sFRP2/3 (Fig. 1e, f), indicating that Wnts participate in the induction of neuronal differentiation of AHPs by factors derived from hippocampal astrocytes⁴.

The interaction of Wnts with their receptors can trigger several signalling pathways, including β -catenin-dependent ('canonical') and β -catenin-independent pathways^{5,8}. The Wnt/ β -catenin signalling reporter Super8XTOPFLASH⁹ was electroporated into AHPs that were subsequently cultured in the presence or absence of a hippocampal astrocyte feeder layer. Co-culture with hippocampal astrocytes resulted in a fourfold increase in luciferase reporter activity (Fig. 1g), suggesting that Wnts derived from hippocampal astrocytes stimulate Wnt/ β -catenin signalling in AHPs. Next, AHPs were electroporated with the Wnt/ β -catenin signalling reporter TOPGAL¹⁰ and cultured on hippocampal astrocytes for 4 d. Expression of the β -galactosidase reporter was found almost exclusively in cells (>95% of the β -gal⁺ cells; $n = 300$) that were positive for DCX (data not shown) or MAP2ab (Fig. 1h), but was not observed in the small fraction of AHPs that differentiated into glial fibrillary acidic protein (GFAP)-positive astrocytes⁴.

Given this apparent association between Wnt signalling and neuronal lineage commitment, we tested the requirement for intact Wnt/ β -catenin signalling in AHPs during neuronal differentiation induced by hippocampal astrocytes. Activation of the canonical Wnt pathway stabilizes β -catenin, which subsequently forms a transcriptionally active complex with members of the TCF/LEF transcription factor family⁵. Mutant forms of TCF/LEF that lack the ability to bind β -catenin act as dominant-negative regulators of Wnt/ β -catenin signalling. AHPs were electroporated with a dominant-negative Lef1 (dnLef1) expression construct or empty vector (see Methods) and plated onto a hippocampal astrocyte feeder layer. After 4 d of co-culture, differentiation into DCX-positive neurons was reduced by 50% in dnLef1-expressing AHPs as compared with controls ($P < 0.05$, Mann-Whitney rank sum test; Fig. 1i). Taken together, these results indicate that astrocyte-derived Wnts and intact Wnt/ β -catenin signalling in AHPs are substantial contributors to the neuronal differentiation of AHPs induced by hippocampal astrocytes.

We sought to determine whether Wnt signalling is sufficient to enhance neurogenesis from AHPs. To this end we focused on the Wnt family member Wnt3, which stimulates Wnt/ β -catenin signalling in AHPs (Supplementary Fig. 2). AHPs were transduced with a retroviral vector expressing Wnt3 under the control of the tetracycline-suppressible *tet* operator¹¹. Wnt3-expressing AHPs in which

¹Laboratory of Genetics, The Salk Institute, La Jolla, California 92037, USA. ²GSF-National Research Centre for Environment and Health, Institute of Developmental Genetics, 85764 Munich, Neuherberg, Germany. ³Institute for Cell Engineering, Departments of Neurology and Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

*These authors contributed equally to this work.

transgene expression was suppressed by the addition of doxycycline (100 ng ml^{-1}) did not differ from wild-type AHPs in generating neurons, astrocytes and oligodendrocytes under differentiating conditions (data not shown). On removal of doxycycline, however, the presence of Wnt3 stimulated a roughly fivefold increase in the percentage of cells that differentiated into neurons as compared with wild-type AHPs (Fig. 2a, b) or Wnt3-expressing AHPs cultured in the presence of doxycycline (data not shown). By contrast, differentiation into oligodendrocytes or astrocytes was not significantly increased (Supplementary Fig. 3), indicating that Wnt3 specifically enhanced the generation of neurons.

Increased neurogenesis in the presence of Wnt3 could be caused by promotion of cell survival, increased proliferation of neuronally committed progenitors or enhanced neuronal cell fate instruction. We determined apoptotic cell death by a TdT-mediated dUTP nick

end labelling (TUNEL) assay every 24 h for 4 d in Wnt3-expressing and wild-type AHP cultures under differentiating conditions. No significant difference in the percentage of cells undergoing apoptotic cell death was observed between the experimental groups, indicating that Wnt3 did not affect survival under these conditions (data not shown). We then evaluated the effects of Wnt3 on neuroblast proliferation. Parallel cultures of Wnt3-expressing and wild-type AHPs were cultured under differentiating conditions and fixed at 24-h intervals over 4 d after a 2-h pulse of $10 \mu\text{M}$ bromodeoxyuridine (BrdU). As expected, the percentage of DCX-positive cells, which comprises proliferating neuroblasts and immature postmitotic neurons⁷, was significantly increased in Wnt3-expressing AHPs as compared with controls at all time points (Supplementary Fig. 4). In the DCX-positive population, the neuroblast fraction was identified by coexpression of the proliferation marker Ki67. The ratio of

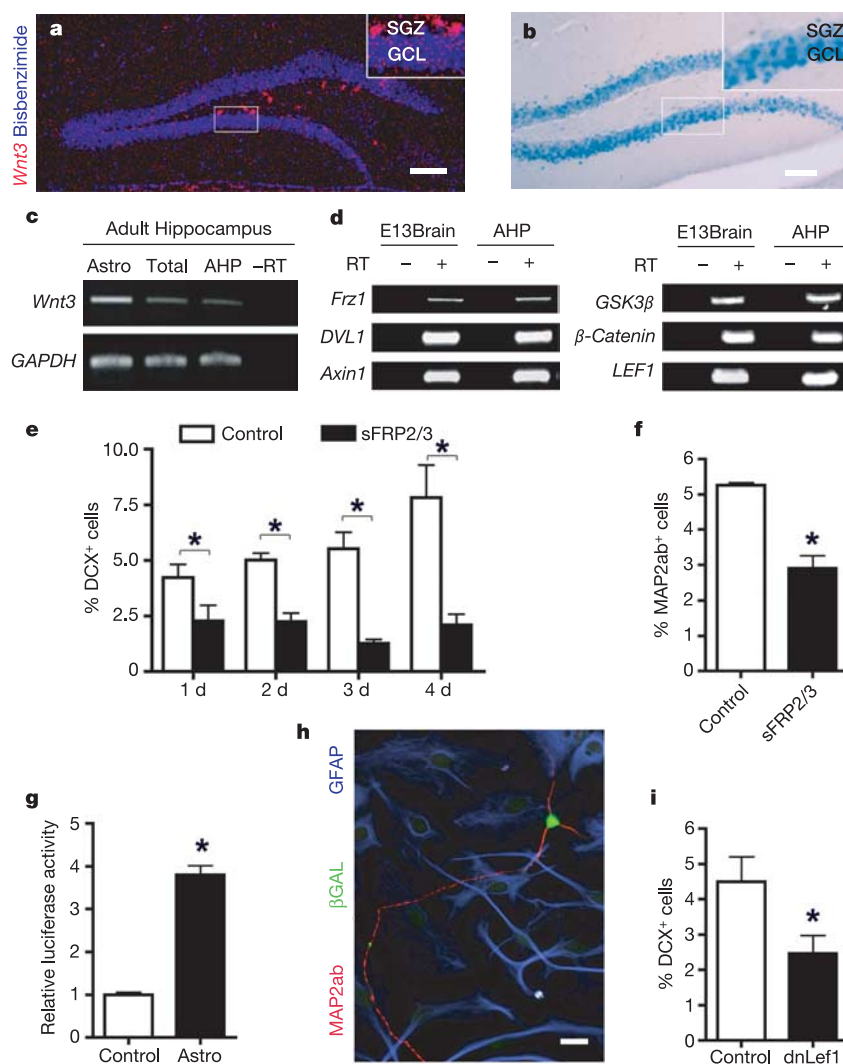


Figure 1 | Wnt signalling is involved in hippocampal-astrocyte-induced neurogenesis from AHPs. **a**, Wnt3 is expressed in close proximity to the SGZ of the adult mouse; Wnt3 mRNA is red, the bisbenzimidide counterstain is blue. Scale bar, $80 \mu\text{m}$. GCL, granule cell layer. **b**, β -Galactosidase staining (blue) in BATGAL mice shows activity of the Wnt/ β -catenin signalling pathway in the SGZ and granule cell layer of the dentate gyrus. **c**, RT-PCR confirms Wnt3 expression in the adult hippocampus (Total) and adult hippocampal astrocytes (Astro). Less Wnt3 mRNA is expressed in AHPs and is not sufficient to stimulate β -catenin signalling significantly (Supplementary Fig. 1). GAPDH mRNA was measured as an internal control. **d**, RT-PCR analysis of Wnt signalling components in AHPs. *Frz1*, *Frizzled1*; *DVL1*, *Dishevelled1*. cDNA from embryonic day 13 brain was

analysed as a positive control. **e**, **f**, Decrease in the percentage of neurons generated from AHPs in co-cultures of AHPs and hippocampal astrocytes caused by the addition of sFRP2/3. **g**, Hippocampal astrocytes stimulate β -catenin signalling in AHPs. As a control, AHPs were cultured in the absence of hippocampal astrocytes. **h**, AHPs expressing MAP2ab (red) show activation of the canonical Wnt signalling pathway (TOPGAL reporter; green) in co-cultures with hippocampal astrocytes (GFAP; blue). Scale bar, $20 \mu\text{m}$. **i**, Blocking canonical Wnt signalling by expression of dnLef1 in AHPs reduces their neuronal differentiation in hippocampal astrocyte co-cultures. Asterisks indicate a statistical difference between experimental groups ($P < 0.05$, Mann-Whitney rank sum test [e, f, i] or Student's *t*-test [g]). Error bars represent the s.e.m.

BrdU-positive neuroblasts ($\text{BrdU}^+\text{DCX}^+\text{Ki67}^+$) per total number of neuroblasts ($\text{DCX}^+\text{Ki67}^+$) was 2.5-fold higher in Wnt3-exposed AHPs on day 2 and remained significantly higher on day 3 (Fig. 2c, d). Despite this enhanced neuroblast proliferation, no change in overall proliferation was observed, as determined by the percentage of BrdU-positive cells among the total population (Fig. 2d), which was most probably due to the relatively small number of neuroblasts (<2% on day 2). The absence of detectable changes in bulk proliferation, however, indicated that Wnt3 does not have strong proliferative effects on cells other than neuroblasts under the current experimental conditions. The specificity of the proliferative effect was further substantiated by our finding that oligodendrocyte precursors did not show increased proliferation in the presence of Wnt3 (Supplementary Fig. 3).

Next, we determined whether Wnt3 also influences the neuronal fate choice of AHPs. To distinguish the proliferative effects of Wnt3 from potential instructive effects, differentiation was studied under growth-arresting conditions. Cells were cultured in differentiating conditions in the continuous presence of the DNA polymerase inhibitor aphidicolin¹² ($10 \mu\text{g ml}^{-1}$) or the ribonucleotide reductase inhibitor hydroxyurea¹³ ($1 \mu\text{M}$). For the first 18 h, Wnt3-expressing AHPs were cultured in the presence of doxycycline to inhibit expression of Wnt3 before the onset of growth arrest. Doxycycline was then removed to allow expression of Wnt3 and the cells were cultured for an additional 40 h. During this period, BrdU was added to label cells that continued to proliferate despite the treatment with aphidicolin or hydroxyurea. Less than 5% of cells incorporated BrdU, indicating that most cells (~95%) were growth-arrested during the

final 40 h of the assay (data not shown). In the growth-arrested population, we observed a 5–10-fold increase in the percentage of DCX-positive neurons generated from Wnt3-expressing AHPs as compared with controls (Fig. 2e, f). These results show that, in addition to stimulating neuroblast proliferation, Wnt3 instructs AHPs to adopt a neuronal fate.

Finally, we tested whether Wnt signalling regulates neurogenesis *in vivo*. Adult BATGAL mice were injected with a single dose of BrdU and killed after 24 h. Wnt/ β -catenin pathway activity was observed in BrdU-positive cells in the SGZ (Fig. 3a) and in DCX-expressing cells with morphological features of proliferating neuroblasts⁷ (Fig. 3b). These *in vivo* results are consistent with our *in vitro* finding that Wnt/ β -catenin signalling regulates neuronal fate commitment and neuroblast proliferation.

For *in vivo* loss- and gain-of-function studies, we used a self-inactivating lentiviral vector (LV)¹⁴ expressing a bicistronic cassette encoding for a Wnt signalling inhibitor or stimulator, an internal ribosomal entry site (IRES) and GFP. To block Wnt signalling *in vivo*, we used a secreted mutant Wnt1 protein (dnWnt), which non-autonomously blocks Wnt signalling *in vivo*^{15,16}. Luciferase assays showed that LV-dnWnt reduced Wnt3-initiated β -catenin signalling in AHPs (Supplementary Fig. 5). In addition, we determined that overexpression of dnWnt in hippocampal astrocytes did not increase cell death in co-cultures of AHPs and hippocampal astrocytes (Supplementary Fig. 6). LV-dnWnt or control LV expressing GFP (LV-GFP) lentiviruses were stereotactically injected into the dentate gyrus of young (8–9 weeks) adult rats ($n = 8$). After a 3-week period to allow transgene expression and to minimize confounding factors

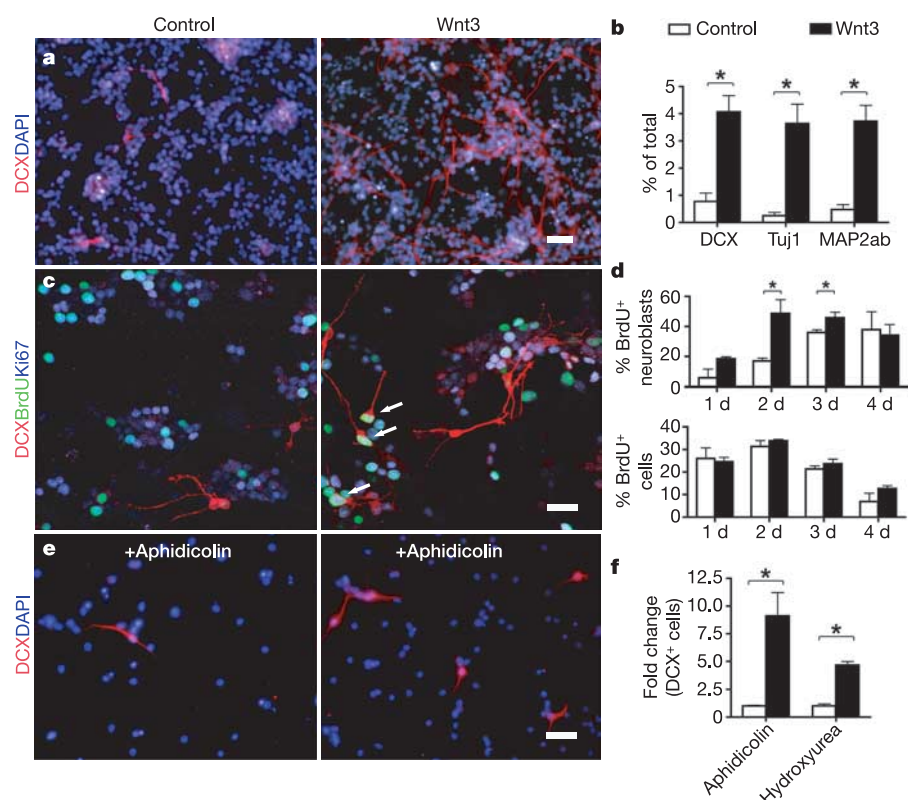


Figure 2 | Wnt3 is sufficient to increase neuronal production from AHP.

a, Overexpression of Wnt3 in AHP leads to more cells expressing the early neuronal marker DCX (red). The DAPI counterstain is in blue.

b, Quantitative analysis of three markers of neurogenesis after 4 d with or without Wnt3 overexpression. **c**, **d**, Analysis of proliferation by BrdU incorporation. Exposure to Wnt3 specifically increases cell division in the neuroblast population, whereas overall BrdU incorporation does not change significantly. Neuroblast proliferation was compared by determining the

percentage of neuroblasts, namely DCX (red) and Ki67 (blue) double-positive cells, that had incorporated BrdU (green; arrows in **c**). **e**, **f**, Blocking cell division by aphidicolin (**e**, **f**) or hydroxyurea (**f**) shows that neuronal fate commitment (**e**; DCX, red) is increased 5–10-fold by Wnt3. The fraction of DCX-expressing cells in Wnt3-expressing AHPs is normalized to the DCX-positive fraction in control cells. Asterisks indicate a statistical difference between experimental groups ($P < 0.05$, Student's *t*-test). Error bars represent the s.e.m. Scale bars, 25 μm .

caused by the surgery, rats were injected with 200 mg per kg (body-weight) of BrdU daily for 7 d to label proliferating AHPs and their progeny. Rats were perfused 24 h after the final BrdU injection. Quantification of TUNEL-positive cells in the transduced dentate gyrus areas did not show any significant difference between the experimental groups (Supplementary Fig. 6). To evaluate potential changes in neurogenesis, we determined the number of BrdU-positive cells and the percentage of BrdU-positive cells that expressed DCX in the targeted (GFP-expressing) areas. As compared with rats injected with LV-GFP, those injected with LV-dnWnt showed a marked reduction in the number of BrdU-positive cells

($11,170 \pm 1,122$ and $2,751 \pm 281.4$ per mm^3 , respectively, $P < 0.0001$, Student's *t*-test; Fig. 3c) and in the percentage of BrdU-positive cells that differentiated into DCX-positive neurons ($57.62 \pm 4.65\%$ and $31.03 \pm 5.48\%$, respectively, $P < 0.005$, Student's *t*-test; Fig. 3d), resulting in roughly an eightfold reduction in neurogenesis in the targeted areas ($6,302 \pm 571.7$ versus 813.6 ± 118.7 new DCX-positive neurons per mm^3 , respectively; Fig. 3e).

To examine the effects of increased Wnt signalling on hippocampal neurogenesis, LV expressing Wnt3 (LV-Wnt3) or control LV-GFP was injected into the dentate gyrus of mature (15-week-old) rats

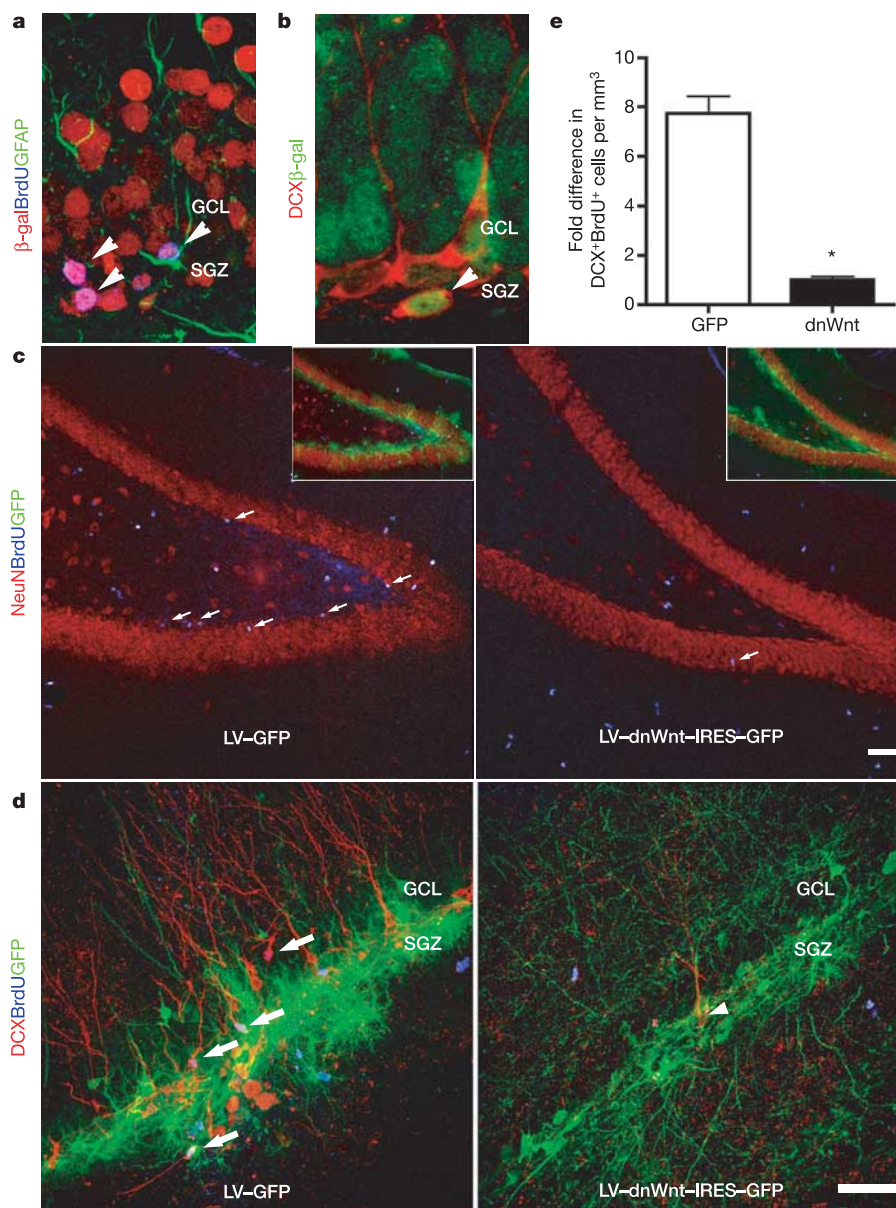


Figure 3 | Blocking Wnt signalling *in vivo* suppresses adult hippocampal neurogenesis. **a**, Wnt/ β -catenin reporter activity (β -gal, red) in BrdU-positive progenitors (blue, arrowheads) in the SGZ of adult BATGAL mice. GFAP, green. **b**, Reporter activity (β -gal, green) in DCX-positive cells (red) with short processes oriented parallel to the dentate granule cell layer (arrowhead). This morphology is consistent with proliferating neuroblasts. **c**, Stereotactic injection of LV-GFP or LV-dnWnt (LV-dnWnt-IRES-GFP) into the adult dentate gyrus. Expression of dnWnt protein inhibits proliferation as determined by BrdU incorporation (blue, arrows) in the SGZ. Inset shows the viral infection, marked by expression of the GFP reporter (green). Mature dentate granule neurons are stained for NeuN

(red). Scale bar, 50 μm . **d**, dnWnt inhibits expression of the neuroblast marker DCX (red, arrowhead). Newborn neurons double-labelled with DCX and BrdU (arrows) are almost eliminated as compared with controls. Scale bar, 70 μm . **e**, Differences in the number of newborn neurons. The rate of neurogenesis was calculated as the number of DCX and BrdU double-positive cells per volume of transduced dentate granule cell layer. The average number of new neurons per transduced volume in dnWnt-injected rats was used for normalization. Asterisks indicate a statistical difference between experimental groups ($P < 0.0001$, Student's *t*-test). Error bars represent the s.e.m.

($n = 6$). Injections of BrdU (100 mg per kg for seven consecutive days) were started 3 weeks after the viral injections, and rats were perfused 24 h after the final BrdU injection. In the transduced areas, rats injected with LV-Wnt3 showed a significantly higher percentage of BrdU-positive cells expressing DCX as compared with those injected with LV-GFP ($70.29 \pm 4.29\%$ and $53.93 \pm 6.76\%$, $P < 0.05$, Mann-Whitney rank sum test). These newborn immature

neurons were often found in large clusters of more than eight DCX and BrdU double-positive cells in rats injected with LV-Wnt3 (Fig. 4a). We also observed a clear trend towards increased numbers of BrdU-positive cells in rats injected with LV-Wnt3 as compared with controls injected with LV-GFP ($9,738.1 \pm 2,460.1$ and $6,211.1 \pm 558.3$ per mm^3 , respectively). Together, these changes led to a doubling in neurogenesis in the targeted areas in rats injected with LV-Wnt3 ($7,010.9 \pm 1,963.7$ versus $3,324.9 \pm 543.0$ new DCX-positive neurons per mm^3 , $P < 0.05$, Mann-Whitney rank sum test; Fig. 4b), showing that increased Wnt signalling is sufficient to stimulate adult hippocampal neurogenesis.

Because genetic manipulation of the Wnt pathway often results in embryonic lethality, little is known about the role of Wnt signalling in the adult central nervous system. Our work identifies Wnt signalling as a regulatory pathway in adult hippocampal neurogenesis involved in the control of neuronal fate commitment and the proliferation of neuronally committed precursor cells. The extent to which the inhibition of Wnt signalling reduces adult hippocampal neurogenesis shows that Wnt signalling has a central role in these processes, although future experiments must address its relative contribution to proliferation and differentiation, as well as its potential involvement in later maturational events. There is evidence that precursor cell responses to Wnts can be modified by other signalling molecules^{17–20}. We expect that such interactions are also involved in the regulation of AHPs given that adult hippocampal neurogenesis can be modulated, albeit to a far lesser extent, by *in vivo* inhibition of several other signalling pathways^{21–23}. Finally, the discovery that hippocampal neurogenesis is almost extinguished by inhibition of Wnt signalling provides a powerful tool to study the as yet elusive role of adult neurogenesis and Wnt signalling in hippocampal function and plasticity.

METHODS

Cell culture. The isolation, characterization and culturing of AHPs used in this study have been described²⁴. Fibroblast growth factor 2 was withdrawn to promote differentiation. The generation of AHPs expressing GFP has been described⁴. AHP cell lines expressing Wnt3 were generated by transduction with the pPIT-Wnt3 vector (Wnt3 cDNA was a gift from R. Nusse, Stanford University, CA). For proliferation studies, $10 \mu\text{M}$ BrdU (Sigma-Aldrich) was added for 2 h before fixation. To inhibit proliferation, we exposed AHPs to either $10 \mu\text{g ml}^{-1}$ of aphidicolin (Sigma) and $200 \mu\text{M}$ thymidine (Sigma) or $1 \mu\text{M}$ hydroxyurea (Sigma).

Primary astrocytes were isolated from rat hippocampus and cultured as described⁴. For co-cultures of AHPs and hippocampal astrocytes, AHPs were plated on a confluent astrocyte feeder layer at a density of 5×10^3 cells per cm^2 in serum-free conditions⁴. Recombinant sFRP2/3 protein (R&D Systems) was added at a concentration of 500 ng ml^{-1} .

Plasmids. We used the following reporter plasmids: Super8xTOPFLASH⁹ (with a TCF/LEF-binding motif) or Super8xFOPFLASH⁹ (mutant motif; a gift from R. Moon, University of Washington, Seattle, WA), TOPGAL (a gift from E. Fuchs, Rockefeller University, NY), and Renilla-Luc under the control of the human elongation factor 1 promoter (internal control). For expression of Wnt3, we used the retroviral vector pPIT (a gift from D. Schaffer, University of California, Berkeley, CA). For expression of dnLef1 (a gift from C. Fryer, Salk Institute, La Jolla, CA), we used the retroviral vector PMY-IRES-GFP²⁵, which contains an IRES-GFP cassette that allows identification of transduced or electroporated cells.

Luciferase assays. AHPs were electroporated with reporter plasmids by a nucleofactor device (Amaxa). R-Luc was co-electroporated as an internal control. Forty-eight hours after electroporation, luciferase activity in $10 \mu\text{l}$ of lysis supernatant was measured with the Dual-Luciferase Reporter Assay System (Promega) and an LB 9501 luminometer (Bechtold).

Immunostaining. Tissue and cultured cells were fixed with 4% paraformaldehyde and processed for immunostaining as described²⁶.

Cell counting. The percentage of antibody-labelled cells was determined by evaluating 1,000 cells in at least five randomly chosen fields of view. To evaluate the proliferation of neuroblasts, we analysed 100–200 DCX-positive cells for expression of Ki67 and incorporation of BrdU. Experiments were done in triplicate. Statistical analysis was done with the nonparametric Mann-Whitney rank sum test or Student's *t*-test.

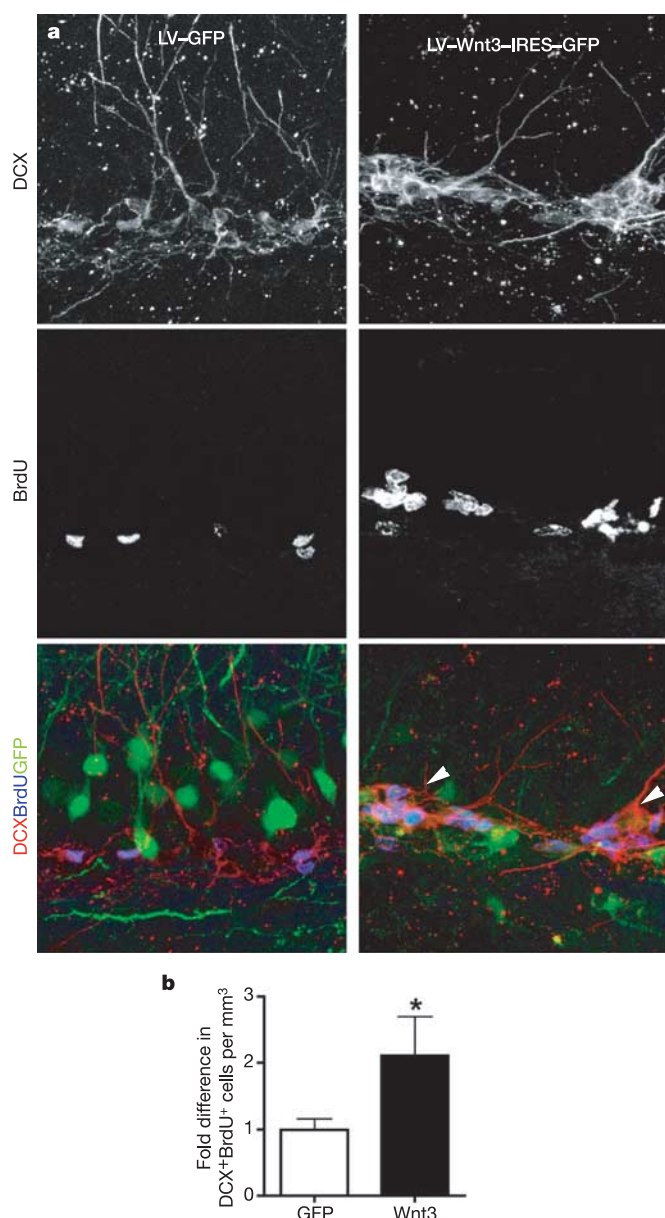


Figure 4 | Enhanced Wnt signalling *in vivo* increases adult hippocampal neurogenesis. **a**, LV-GFP or LV-Wnt3 was stereotactically injected into the adult rat dentate gyrus. Newborn neurons double labelled for DCX (red) and BrdU (blue) were found mostly in large clusters of about 8–10 cells (arrowheads) in the SGZ of rats injected with LV-Wnt3. Newborn neurons in rats injected with LV-GFP were predominantly found in smaller clusters of 2–4 cells. GCL, granule cell layer. Scale bar, $35 \mu\text{m}$. **b**, Differences in the number of newborn neurons in control GFP and Wnt3-expressing hippocampi. The rate of neurogenesis in the experimental groups was calculated as the number of DCX and BrdU double-positive cells per volume of infected dentate granule cell layer. The average number of new neurons per infected volume in GFP-injected rats was used for normalization. Asterisk indicates a statistical difference between experimental groups ($P < 0.05$, Mann-Whitney rank sum test). Error bars represent the s.e.m.

In situ hybridization. *In situ* hybridization was done according to standard protocols²⁷. Slides were subsequently processed for emulsion autoradiography using Kodak NTB-2 emulsion (Eastman Kodak). Sense controls yielded only nonspecific background labelling (data not shown).

Lentiviral vectors. The control vectors were CSC.cPPT.hCMV.GFP.Wpre (loss-of-function studies) and pRRL.SIN.cPPT.hPGK.GFP (gain-of-function studies)¹⁴. To generate the dnWnt-IRES-GFP vector, we cloned the cDNA for dnWnt upstream of the IRES and GFP and inserted the bicistronic cassette in place of the GFP sequence in the CSC.cPPT.hCMV.GFP.Wpre vector. To construct the Wnt3-IRES-GFP vector, we cloned the cDNA of Wnt3 upstream of the IRES and GFP and inserted the bicistronic cassette in place of the GFP sequence in the pRRL.SIN.cPPT.hPGK.GFP.Wpre vector. Concentrated LV stocks, pseudotyped by the vesicular stomatitis viral envelope, were produced as described¹⁴. Expression titres, determined on HeLa cells by FACS analysis, were 5×10^9 to 1×10^{10} transducing units per ml with an HIV-1 p24 concentration of 200–400 $\mu\text{g ml}^{-1}$.

In vivo neurogenesis experiments. All rat procedures were done in accordance with protocols approved by the animal care and use committee of The Salk Institute for Biological Studies. For loss-of-function studies, 1.5 μl of vector concentrate were stereotactically injected into the right hippocampal dentate gyrus (AP -4 , ML ± 2 , DV -3.5 from Bregma, nose piece -3.3) of adult female Fisher 344 rats aged 8–9 weeks ($n = 8$ per group). For gain-of-function studies, 1.5 μl of vector concentrate were stereotactically injected into the right hippocampal dentate gyrus of adult female Fisher 344 rats aged 15 weeks ($n = 6$ per group).

Stereology. Areas infected by LVs were identified by expression of GFP. Experimental groups showed comparable viral spread, as determined by the presence of GFP-expressing cells along the anterior–posterior axis of the dentate gyrus (720–960 μm). We counted BrdU-positive cells in the dentate granule cell layer in a one-in-six series of sections (240- μm apart) throughout the rostro-caudal extent of the infected dentate granule cell layer. NeuN immunoreactivity was used to measure the granule cell layer volume. The granule cell area was traced by using a StereoInvestigator semiautomatic stereology system (MicroBrightfield) and a $10\times$ objective. The proliferation rate was expressed as BrdU cells per volume of dentate granule cell layer²⁶.

Phenotype analysis. Sections (40- μm thick) containing dentate gyrus areas infected by LVs were identified by expression of GFP. All BrdU-positive cells in the subgranular and granular layer of the hippocampal dentate gyrus were analysed by confocal microscopy for labelling for both BrdU and DCX. The differentiation rate was determined by dividing the number of BrdU-positive cells that stained for DCX by the number of BrdU-positive cells evaluated.

Received 3 June; accepted 22 July 2005.

- Alvarez-Buylla, A. & Lim, D. A. For the long run: maintaining germinal niches in the adult brain. *Neuron* **41**, 683–686 (2004).
- Kleber, M. & Sommer, L. Wnt signalling and the regulation of stem cell function. *Curr. Opin. Cell Biol.* **16**, 681–687 (2004).
- Maretto, S. *et al.* Mapping Wnt/ β -catenin signalling during mouse development and in colorectal tumors. *Proc. Natl Acad. Sci. USA* **100**, 3299–3304 (2003).
- Song, H., Stevens, C. F. & Gage, F. H. Astroglia induce neurogenesis from adult neural stem cells. *Nature* **417**, 39–44 (2002).
- Moon, R. T. Canonical Wnt/ β -catenin Signaling. *Sci. STKE* **2004**, TR5 (2004).
- Rattner, A. *et al.* A family of secreted proteins contains homology to the cysteine-rich ligand-binding domain of frizzled receptors. *Proc. Natl Acad. Sci. USA* **94**, 2859–2863 (1997).
- Couillard-Despres, S. *et al.* Doublecortin expression levels in adult brain reflect neurogenesis. *Eur. J. Neurosci.* **21**, 1–14 (2005).
- Veeman, M. T., Axelrod, J. D. & Moon, R. T. A second canon. Functions and mechanisms of β -catenin-independent Wnt signalling. *Dev. Cell* **5**, 367–377 (2003).
- Veeman, M. T., Slusarski, D. C., Kaykas, A., Louie, S. H. & Moon, R. T. Zebrafish prickles, a modulator of noncanonical Wnt/Fz signalling, regulates gastrulation movements. *Curr. Biol.* **13**, 680–685 (2003).
- DasGupta, R. & Fuchs, E. Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation. *Development* **126**, 4557–4568 (1999).
- Sakurada, K., Ohshima-Sakurada, M., Palmer, T. D. & Gage, F. H. Nurr1, an orphan nuclear receptor, is a transcriptional activator of endogenous tyrosine hydroxylase in neural progenitor cells derived from the adult brain. *Development* **126**, 4017–4026 (1999).
- McBeath, R., Pirone, D. M., Nelson, C. M., Bhadriraju, K. & Chen, C. S. Cell shape, cytoskeletal tension, and RhoA regulate stem cell lineage commitment. *Dev. Cell* **6**, 483–495 (2004).
- Schwartz, K. A., Lanciloti, N. J., Moore, M. K., Campione, A. L. & Chandar, N. p53 transactivity during *in vitro* osteoblast differentiation in a rat osteosarcoma cell line. *Mol. Carcinog.* **25**, 132–138 (1999).
- Consiglio, A. *et al.* Robust *in vivo* gene transfer into adult mammalian neural stem cells by lentiviral vectors. *Proc. Natl Acad. Sci. USA* **101**, 14835–14840 (2004).
- Garcia-Castro, M. I., Marcelle, C. & Bronner-Fraser, M. Ectodermal Wnt function as a neural crest inducer. *Science* **297**, 848–851 (2002).
- Hoppler, S., Brown, J. D. & Moon, R. T. Expression of a dominant-negative Wnt blocks induction of MyoD in *Xenopus* embryos. *Genes Dev.* **10**, 2805–2817 (1996).
- Israsena, N., Hu, M., Fu, W., Kan, L. & Kessler, J. A. The presence of FGF2 signalling determines whether β -catenin exerts effects on proliferation or neuronal differentiation of neural stem cells. *Dev. Biol.* **268**, 220–231 (2004).
- Jamora, C., DasGupta, R., Kocieniewski, P. & Fuchs, E. Links between signal transduction, transcription and adhesion in epithelial bud development. *Nature* **422**, 317–322 (2003).
- Duncan, A. W. *et al.* Integration of Notch and Wnt signalling in hematopoietic stem cell maintenance. *Nature Immunol.* **6**, 314–322 (2005).
- Kleber, M. *et al.* Neural crest stem cell maintenance by combinatorial Wnt and BMP signalling. *J. Cell Biol.* **169**, 309–320 (2005).
- Lai, K., Kaspar, B. K., Gage, F. H. & Schaffer, D. V. Sonic hedgehog regulates adult neural progenitor proliferation *in vitro* and *in vivo*. *Nature Neurosci.* **6**, 21–27 (2003).
- Lee, J., Duan, W. & Mattson, M. P. Evidence that brain-derived neurotrophic factor is required for basal neurogenesis and mediates, in part, the enhancement of neurogenesis by dietary restriction in the hippocampus of adult mice. *J. Neurochem.* **82**, 1367–1375 (2002).
- Cao, L. *et al.* VEGF links hippocampal activity with neurogenesis, learning and memory. *Nature Genet.* **36**, 827–835 (2004).
- Palmer, T. D., Takahashi, J. & Gage, F. H. The adult rat hippocampus contains primordial neural stem cells. *Mol. Cell. Neurosci.* **8**, 389–404 (1997).
- Kitamura, T. *et al.* Retrovirus-mediated gene transfer and expression cloning: powerful tools in functional genomics. *Exp. Hematol.* **31**, 1007–1014 (2003).
- van Praag, H., Kempermann, G. & Gage, F. H. Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neurosci.* **2**, 266–270 (1999).
- Lein, E. S., Zhao, X. & Gage, F. H. Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput *in situ* hybridization. *J. Neurosci.* **24**, 3879–3889 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Clevers, R. Nusse, R.T. Moon, E. Fuchs, R. Marr and J. Nathans for cDNA and plasmids; S. Piccolo and B. Sosa-Pineda for BATGAL reporter mice; R. Summers, K. Nakashima, T. Kuwabara and H. Suh for discussions and suggestions; M.L. Gage for editorial comments; and L. Kitabayashi for technical assistance. D.C.L. and S.J. are supported in part by the Deutsche Forschungsgemeinschaft (DFG, Germany), S.A.C. was supported by NRSA, L.D. was supported by the Association for Medical Research (FRM, France), A.C. is a recipient of a Telethon postdoctoral fellowship. Additional support was provided by grants from the National Institute of Neurological Disorders and Stroke (NINDS) and National Institute on Aging (NIA), the Max Planck Research Award Program funded by the German Ministry for Education, Science, Research, and Technology, the Lookout Fund, the Christopher Reeves Paralysis Foundation, Defense Advanced Research Projects Agency and Project ALS (to F.H.G.), and from the NINDS and NIA (to H.J.S.).

Author Contributions D.C.L. is the leading author. He contributed to the concept, designed and did the experiments, analysed the data, wrote the paper and provided partial financial support. S.A.C. contributed to the concept, designed and did the *in vivo* loss-of-function experiment, contributed reagents and revised the paper. H.J.S. contributed to the concept and the co-culture experiments. L.D. initiated the study, contributed to the concept and did part of the RT-PCR analysis. H.M. contributed to the co-culture experiments. A.C. generated the LV vectors. E.S.L. did the *in situ* hybridization. S.J. analysed the BATGAL mice. H.L. and A.R.D. provided technical support and contributed to analysis of the data. F.H.G. is the senior author. He contributed to the concept, analysed the data, revised the manuscript and provided the main financial support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.H.G. (gage@salk.edu) or D.C.L. (chichung.lie@gsf.de).

LETTERS

Regulation of Lethal giant larvae by Dishevelled

Gretchen L. Dollar¹, Ursula Weber¹, Marek Mlodzik¹ & Sergei Y. Sokol¹

The establishment of polarity in many cell types depends on Lgl, the tumour suppressor product of *lethal giant larvae*, which is involved in basolateral protein targeting^{1–4}. The conserved complex of Par3, Par6 and atypical protein kinase C^{5–8} phosphorylates and inactivates Lgl at the apical surface; however, the signalling mechanisms that coordinate cell polarization in development are not well defined. Here we show that a vertebrate homologue of Lgl associates with Dishevelled, an essential mediator of Wnt signalling, and that Dishevelled regulates the localization of Lgl in *Xenopus* ectoderm and *Drosophila* follicular epithelium. We show that both Lgl and Dsh are required for normal apical–basal polarity of *Xenopus* ectodermal cells. In addition, we show that the Wnt receptor Frizzled 8, but not Frizzled 7, causes Lgl to dissociate from the cortex with the concomitant loss of its activity *in vivo*. These findings suggest a molecular basis for the regulation of cell polarity by Frizzled and Dishevelled.

Polarity is a fundamental property of cells in multicellular organisms that is required for asymmetric division, changes in cell shape,

adhesion and migratory behaviour. Recent studies have indicated that the core mechanisms of cell polarization are conserved^{11,3,5,6,9}. In *Drosophila* epithelial cells, sensory organ precursors and neuroblasts, as well as in mammalian cells, the apical Par (partitioning defective) complex, which consists of the PDZ-domain-containing proteins Par6 and Par3, and atypical protein kinase C (aPKC), regulates apical–basal polarity and proper localization of cell fate determinants^{5–12}. Activation of aPKC in the Par6 complex results in the phosphorylation of Lgl and its dissociation from the cortex^{12–14}. In epithelial cells lacking Lgl, apical markers mislocalize to the basolateral cortex^{2,3,15}. In neuronal precursors, Lgl is required for the asymmetric targeting of cell fate determinants^{13,16}. Lgl binds Syntaxin 4, a component of the basolateral exocytic machinery⁴, and the yeast Lgl homologues Sro7 and Sro77 are required for polarized exocytosis¹⁷. These findings support the view that Lgl controls cell polarity through basolateral targeting.

In developing tissues, individual cells, although capable of intrinsic polarization¹⁸, must polarize correctly in the context of their

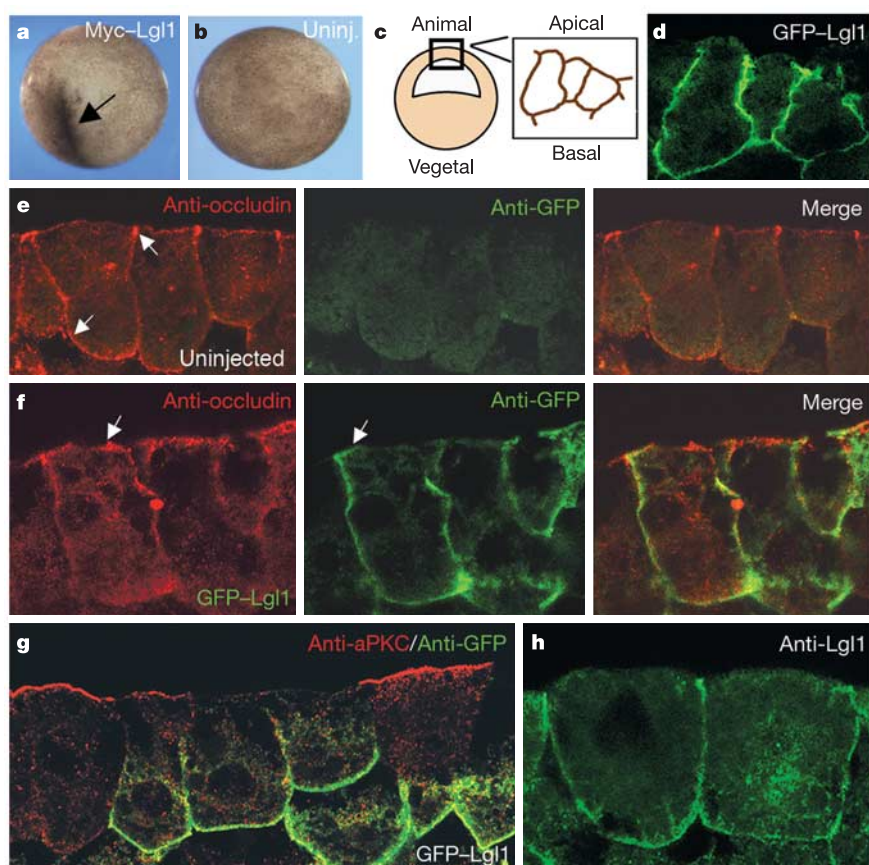


Figure 1 | Lgl1 modulates apical-basal polarity in superficial ectoderm. Two blastomeres of four-cell embryos were injected with 0.5 ng of *Myc-Lgl1* or *GFP-Lgl1* RNA as indicated. **a**, Altered pigmentation in an embryo expressing *Myc-Lgl1*. **b**, Uninjected embryo. **c**, Experimental scheme. **d–g**, Sectioned superficial ectoderm from embryos injected with *GFP-Lgl1* and double-stained with antibodies against GFP (**d**, **e** and **f**, middle; **g**; green) and either occludin (**e** and **f**, left) or aPKC (**g**; red). Apical is at the top. **d**, Basolateral distribution of GFP-Lgl1. **e**, Occludin localization at basolateral surfaces and tight junctions (arrows) in uninjected cells. **f**, In cells expressing GFP-Lgl1 within the same embryo, occludin is ectopically localized to the apical surface (arrow). Right panels in **e** and **f** show merged images. **g**, Loss of apical aPKC (red) in cells expressing Lgl1. **h**, Basolateral localization of endogenous Lgl1.

¹Department of Molecular, Cell and Developmental Biology, Box 1020, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, New York 10029, USA.

environment by responding to extracellular polarizing cues¹. In fact, both coordinated polarization of cells in the plane of epithelial tissue^{19,20} and the alignment of intercalating cells during vertebrate gastrulation^{21,22} require Frizzled (Fz) and Dishevelled (Dsh), which are components of the Wnt signalling pathway²³. In addition, members of the Wnt pathway have been implicated in many processes that also involve the Par6–Par3–aPKC pathway, including asymmetric division of *Drosophila* sensory organ precursors⁷ and mitotic spindle orientation in *Caenorhabditis elegans* blastomeres¹¹. Despite these functional similarities, our understanding of the biochemical communication between these two pathways remains limited.

We identified a *Xenopus* homologue of Lgl as a protein that interacts with Dsh in a yeast two-hybrid screen, raising the possibility that Lgl is regulated by the Wnt pathway. To assay Lgl activity, we determined the effect of overexpressing Lgl in *Xenopus* embryonic ectoderm. Animal pole blastomeres express Lgl (data not shown) and have a well-defined apical–basal polarity, including the apical restriction of aPKC, a negative regulator of Lgl, suggesting that they are an appropriate *in vivo* system in which to examine Lgl activity. We obtained a full-length cDNA encoding *Xenopus* Lgl (see Methods) and expressed it as a fusion protein with green fluorescent protein (GFP) or Myc in *Xenopus* embryos. Just before gastrulation, we observed a marked change in pigment distribution in superficial ectoderm cells overexpressing Lgl (Fig. 1a), as compared with the homogeneous pigmentation of uninjected embryos (Fig. 1b). This phenotype was observed in most embryos injected with either GFP–Lgl or Myc–Lgl RNA ($n > 100$) and was dependent on dose, because the pigment redistribution became more evident when larger amounts of RNA were injected (data not shown).

To see whether these pigment changes reflected cell polarity defects, embryos expressing GFP–Lgl were cross-sectioned and double immunostained for GFP and either aPKC, an apical marker, or occludin, which marks the basolateral membrane and developing tight junctions in stage 10 embryos²⁴ (Fig. 1c–g). Cells expressing GFP–Lgl showed a marked change in polarity as compared with uninjected cells, as judged by the loss of apical aPKC and the spread of occludin to the apical surface. Most GFP–Lgl was localized to the basolateral membrane (Fig. 1c, d), consistent with conserved regulation by endogenous aPKC^{12–14}. Some GFP–Lgl was detected at the apical surface (Fig. 1f, middle), however, suggesting that the observed effects were due to ectopic Lgl activity.

To determine the localization of endogenous Lgl in these cells, we raised antibodies against *Xenopus* Lgl (Supplementary Fig. 1), which showed that it was restricted to the basolateral membrane (Fig. 1h). In addition, an Lgl construct in which conserved aPKC phosphorylation sites were mutated localized mostly to the apical surface and caused identical pigmentation and polarity changes when smaller amounts of RNA were injected (data not shown). These and other observations²⁵ indicate that the overexpression phenotype may result from ectopic Lgl activity at the apical surface. This interpretation is consistent with the ability of a mouse homologue of Lgl to inhibit tight junction formation in tissue culture cells¹⁴.

To confirm that Lgl and Dsh interact *in vivo* in *Xenopus* embryonic cells, immunoprecipitations were carried out with lysates from embryos expressing tagged proteins. The fragment of Lgl identified in the yeast screen (HA–LglC) specifically co-precipitated with full-length Myc–Xdsh (Supplementary Fig. 2a, b). Similarly, in a complementary experiment full-length GFP–Lgl co-precipitated with a region of Xdsh containing the DIX domain (Xdsh-N; Supplementary Fig. 2a, c). In addition, endogenous Dsh associated with both Lgl constructs (Supplementary Fig. 2d, e). Because most Lgl1 and Dsh molecules do not colocalize at their steady-state levels in the cell, the two proteins are likely to interact transiently, reminiscent of the regulation of Lgl by Par6–aPKC.

To evaluate whether Dsh is required for Lgl function, we examined

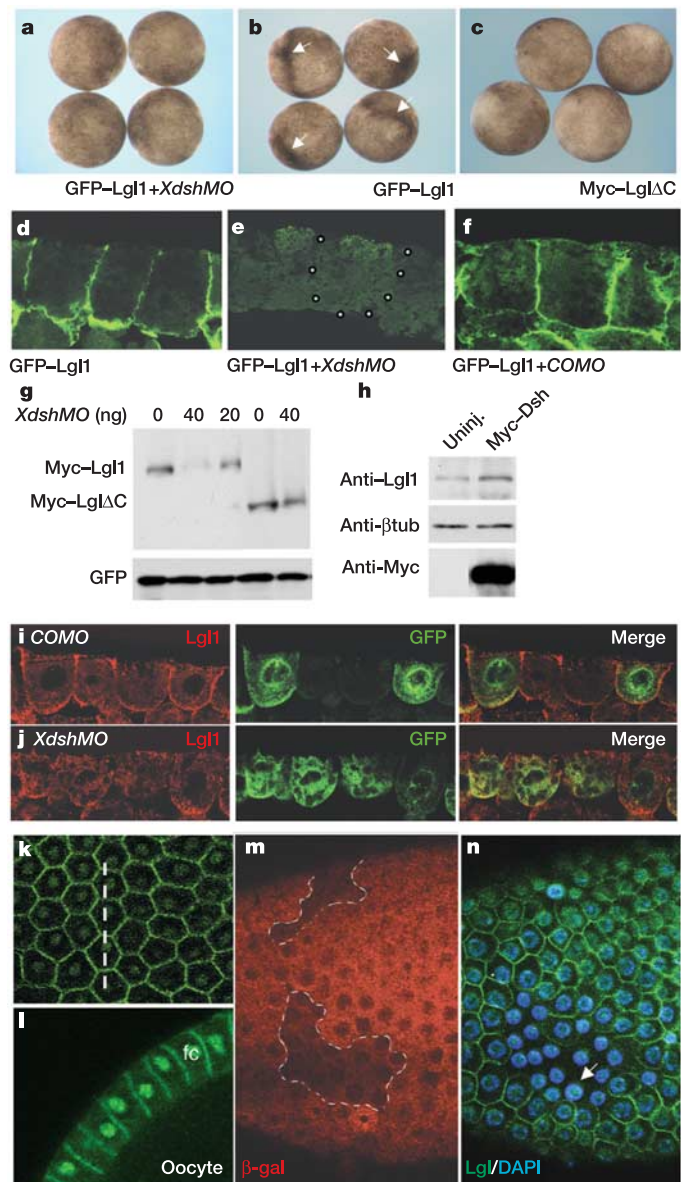


Figure 2 | Dsh is required for Lgl activity and membrane localization.

a, b, Embryos depleted of Dsh with 40 ng of *XdshMO* (**a**) do not show the ectopic pigmentation characteristic of wild-type embryos expressing 0.5 ng of *GFP-Lgl1* RNA (**b**, arrows). **c**, Lgl1 lacking the Dsh-interacting region (Myc–LglΔC) is inactive. **d–f**, Localization of GFP–Lgl1 in superficial ectoderm in the absence (**d**), or presence of *XdshMO* (**e**) or *COMO* (**f**), detected by immunostaining of embryonic sections for GFP. **g**, Myc–Lgl1, but not co-injected GFP, is reduced in embryos previously injected with *XdshMO*. Myc–LglΔC was not significantly altered by *XdshMO*. **h**, Myc–Dsh RNA (1 ng) increases the amount of endogenous Lgl1. β-Tubulin shows equal loading. **i, j**, *XdshMO* (**j**), but not *COMO* (**i**), disrupts basolateral distribution of endogenous Lgl1. Injected cells were identified by the co-injection of GFP RNA (**i** and **j**, middle); sections were double-stained with antibodies against LglC (red) and GFP (green). Right panels in **i** and **j** show merged images. **k, l**, *Drosophila* Lgl protein is found at the lateral membrane and in the nucleus of the follicular epithelium. **k**, Lateral view; **l**, optical cross-section through the follicle cell (fc) layer as indicated in **k**. **m, n**, Ovaries containing *dsh*-null follicle cell clones were co-stained with antibodies against β-gal (**m**, red), Lgl (**n**, green) and DAPI (blue) to show tissue integrity. *Dsh* mutant clones are identified by the absence of β-gal, clone borders are indicated (**m**). Lgl is delocalized from the cell membrane in the clones (**n**, arrow). This effect was observed in 38% of large *dsh* mutant clones ($n = 13$). Posterior is to the left.

the activity of GFP-Lgl1 in embryos injected with a morpholino antisense oligonucleotide (*XdshMO*) that has been shown to reduce specifically endogenous *Xdsh*²⁶. Depletion of *Xdsh* completely suppressed ectopic pigmentation caused by *Lgl1* RNA, indicating that Dsh is required for Lgl1 activity (Fig. 2a, b). Consistent with this, an Lgl1 construct lacking the region involved in Dsh binding did not cause the pigmentation changes characteristic of full-length Lgl1 (Fig. 2c), although it was expressed in similar amounts (Fig. 2g, and data not shown).

We examined whether Dsh is required for the subcellular localization of Lgl. The membrane localization of GFP-Lgl1 was abolished in embryos injected with *XdshMO*, but not in control embryos (Fig. 2d–f). Western blot analysis showed that Myc-Lgl1 protein was decreased in embryos depleted of Dsh in a dose-dependent manner (Fig. 2g). This decrease was specific to Lgl1, because *XdshMO* did not alter the amount of GFP, which was coexpressed in the same embryos as an internal control. In addition, *XdshMO* did not have the same effect on the expression of Myc-Lgl Δ C (Fig. 2g). In a complementary experiment, overexpression of *Dsh* RNA in blastula ectoderm resulted in an increase in endogenous Lgl1 (Fig. 2h). These observations indicate that Dsh may be involved in regulating the localization and stability of Lgl. We confirmed that Dsh is required for localization of Lgl1 to the membrane by examining the distribution of endogenous Lgl1 in *XdshMO*-injected cells, using GFP as a lineage tracer. Almost no basolateral Lgl1 was detected in GFP-positive cells that received *XdshMO*, whereas Lgl1 was unaffected in cells injected with a control morpholino (*COMO*; Fig. 2i, j). We conclude that Dsh is required for Lgl1 localization and activity *in vivo*, possibly by stabilizing Lgl.

To assess whether the regulation of Lgl by Dsh is conserved in other organisms, we examined the distribution of Lgl in *dsh*-null mutant clones in *Drosophila*. To avoid a possible compensatory effect of maternal Dsh, we tested the requirement for Dsh in the follicle cell epithelium, which forms around the oocyte relatively late in development. In follicle cells, Lgl is localized to the lateral membrane and in the nucleus (Fig. 2k, l). Mutant *dsh* follicle cell clones were generated with the FLP/FRT system²⁷. We found that Lgl was delocalized from

the membrane of *dsh* mutant cells (Fig. 2m, n). These results extend the observations made in *Xenopus* ectoderm to the fly follicular epithelium and show that Dsh has a conserved role in subcellular localization of Lgl.

We next tested whether Lgl and Dsh are required for apical–basal polarity in embryonic ectoderm, because our findings predicted that depletion of Lgl1 and Dsh would cause similar apical–basal polarity defects. A morpholino oligonucleotide was designed that specifically decreased Lgl1 protein *in vivo* by gastrulation stages (*LglMO*, Supplementary Fig. 3). Embryos injected with *LglMO*, *XdshMO* or *COMO*, together with GFP RNA as a lineage tracer, were sectioned and stained with antibodies to aPKC, occludin, β -catenin and β 1-integrin, along with antibodies against GFP. Both Lgl1 and Dsh depletion resulted in a loss of apical aPKC (Fig. 3b, c) and in ectopic accumulation of occludin at the apical surface (Fig. 3e, f), as compared with cells injected with *COMO* (Fig. 3a, d). By contrast, the basolateral localization of β -catenin and β 1-integrin was unaffected by Lgl and Dsh depletion (Fig. 3g–l). These results show that Lgl1 is required for some, but not all, aspects of apical–basal polarity of the ectoderm. The similar defects seen in Dsh-depleted cells are probably due to the role of Dsh in Lgl regulation. Most importantly, these findings establish that Dsh, as well as Lgl1, is involved in apical–basal epithelial polarization.

Because Dsh is an essential mediator of Fz signalling, we examined whether Fz signals can influence Lgl activity. Overexpression of Fz receptors has been shown to mimic aspects of active signalling²⁸. *Fz8*, *Fz7* and *Fz3* RNAs were co-injected with Myc-Lgl1 or GFP-Lgl1 RNA into animal blastomeres of four-cell *Xenopus* embryos. We found that *Fz8*, but not *Fz7* or *Fz3*, RNA completely suppressed Lgl1-dependent pigment redistribution (Fig. 4a, b, and data not shown), indicating that *Fz8* specifically interferes with Lgl function. Moreover, the membrane localization of GFP-Lgl1 was markedly altered by co-expression of *Fz8* (Fig. 4d, e), but not by *Fz7* (Fig. 4f). All three *Fz* RNAs were equally efficient in interfering with morphogenetic movements in later development, consistent with similar amounts of protein expression (data not shown). Overall amounts of Myc-Lgl1 were slightly reduced by coexpression of *Fz8*, but not *Fz7*

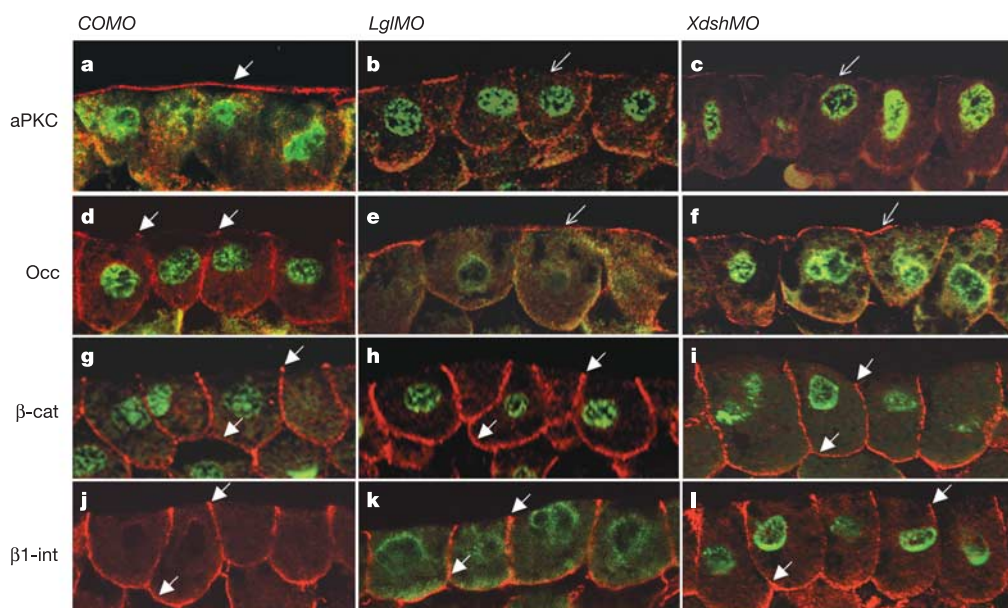


Figure 3 | Lgl1 and Dsh are required for apical–basal polarity in embryonic ectoderm. Embryos injected with *LglMO* (b, e, h, k), *XdshMO* (c, f, i, l) or *COMO* (a, d, g, j), as indicated, were sectioned and immunostained for aPKC (a–c), occludin (d–f), β -catenin (g–i) and β 1-integrin (j–l, red). Injected cells were identified by co-injected GFP (green). Depletion of Lgl1 and Dsh results

in the loss of apical aPKC (b, c) and ectopic occludin (e, f) at the apical surface (open arrows). Basolateral and tight junctional localization of β -catenin (g–i) and β 1-integrin (j–l) was unaffected by *LglMO* or *XdshMO* (filled arrows). The distribution of all markers in cells injected with *COMO* (a, d, g) was indistinguishable from wild type. j, Uninjected embryo.

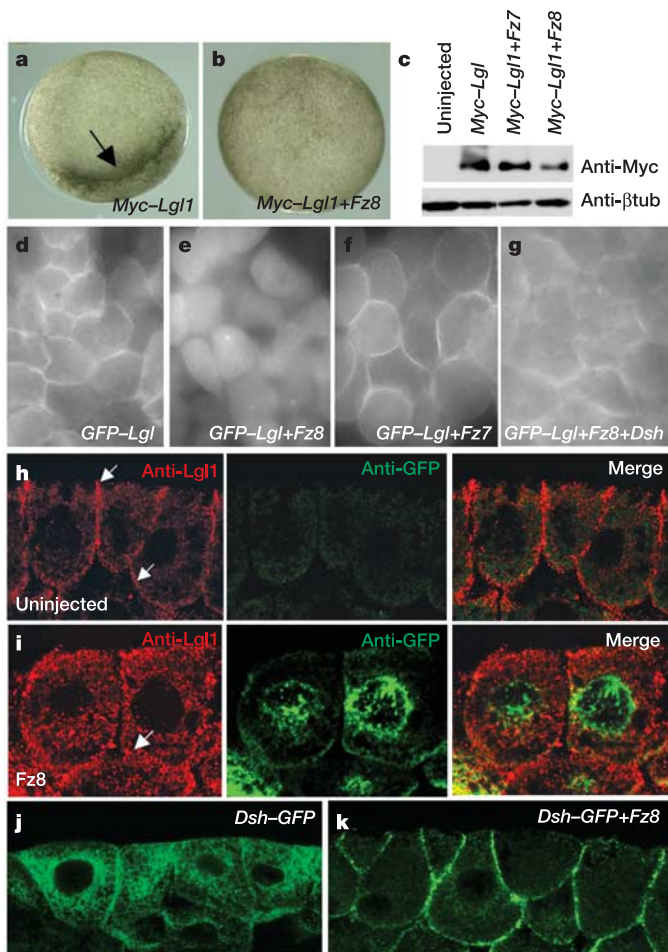


Figure 4 | Fz8 alters localization and activity of Lgl. **a, b,** Changes in pigmentation (arrow) induced by 0.5 ng of *Myc-Lgl1* RNA (**a**) are suppressed by co-injected *Fz8* RNA (**b**). **c,** *Myc-Lgl1* protein in embryos injected with 1 ng of *Fz7* or *Fz8* RNA. β -Tubulin is a loading control. **d–f,** Membrane localization of GFP-Lgl (**d**) is disrupted in embryos co-injected with *Fz8* RNA (**e**), but not *Fz7* RNA (**f**). **g,** Partial rescue of membrane localization of Lgl by 1 ng of *Dsh* RNA. **h–i,** Distribution of endogenous Lgl (red) in uninjected (**h**) and *Fz8*-injected (**i**) ectoderm cells in the same embryo. *Fz8*-injected cells were identified by co-injected GFP (**h** and **i**, middle). Right panels in **h** and **i** show merged images. **j, k,** *Dsh*-GFP is recruited specifically to the basolateral membrane by *Fz8*.

(Fig. 4c). In contrast to *Dsh*-depleted embryos, however, *Myc-Lgl1* was still highly expressed in these embryos, indicating that *Fz8* primarily affects Lgl localization.

Lastly, the delocalization of Lgl in response to *Fz8* was not limited to overexpressed Lgl, but was also observed for endogenous Lgl (Fig. 4h, i). Different Fz receptors seem to function in a compartmentalized manner in the cell, as suggested by their differential distribution along the apical–basal axis of *Drosophila* epithelial cells²⁸. This raises the possibility that in our assay *Fz8* is acting in a dominant-negative manner by sequestering *Dsh* from its required subcellular localization. We found that *Fz8*, as well as *Fz7* and *Fz3*, recruits *Dsh*-GFP to the basolateral surface (Fig. 4j, k, and data not shown), however, supporting the notion that the effect of *Fz8* on Lgl is receptor-specific. We conclude that individual Fz receptors can differentially influence the intracellular distribution and activity of Lgl.

The control of Lgl by *Dsh* suggests a general mechanism for the coordinated regulation of cell polarity by extracellular signalling that is based on existing apical–basal polarity cues. For example, localized

Fz–*Dsh* signalling could provide positional cues for the Par3–Par6–aPKC complex by locally depleting Lgl, which acts antagonistically in the formation of this complex^{2,3,14,15}. Consistent with this model, in *Drosophila* embryos *Fz* and *Dsh* colocalize with the apical Par complex in ectoderm cells and restrict the Par complex to the posterior cortex in sensory organ precursors²⁹. A partial disruption of apical–basal polarity in the plane of epithelial tissue by localized *Fz*–*Dsh* signalling could contribute to planar cell polarity, whereas a more complete disruption could result in epithelial–mesenchymal transformation and altered cell movements that are crucial for morphogenesis. Further studies are warranted to examine the interactions between the molecules involved in regulating planar and apical–basal polarity and those controlling morphogenetic movements in development.

METHODS

DNA constructs. A cDNA encoding a protein fragment with significant sequence similarity to Lgl was obtained in a yeast two-hybrid screen for *Xenopus* gastrula proteins that interact with the amino-terminal domain of *Xdsh* (amino acids 1–134; data not shown). Database searches using this sequence resulted in identification of an expressed sequence tag clone (GenBank clone ID: 6637692) encoding a homologue of *Xenopus* Lgl. We confirmed that the cDNA insert encoded the full-length protein by sequencing (*Xenopus* Lgl1 accession number: AY705978). *Myc-Lgl1* and *GFP-Lgl1* were constructed by inserting the complete coding sequence of *Xenopus* Lgl1 into pXT7–Myc and pCTX–eGFP. *Myc-Lgl1ΔC* was created by subcloning the internal *EcoRI*–*NsiI* fragment into pCTX–Myc. We generated HA–LglC by subcloning the 1.7-kb *EcoRI*–*XhoI* cDNA fragment identified in the yeast two-hybrid screen into pCTX–HA. pCTX vectors were derived from pCS2 by swapping the SP6 and T7 promoters and inserting a multiple cloning site (S. Sokol, unpublished data). Details of cloning are available from the authors on request.

Embryo culture and microinjections. Capped synthetic RNAs were generated by *in vitro* transcription with Sp6 and T7 RNA polymerases using the mMessage mMachine kit (Ambion). Embryos were obtained from *Xenopus* females and cultured in 0.1× Marc's modified Ringer's medium (MMR) as described³⁰. For microinjection, embryos were transferred to 3% Ficoll 400 (Pharmacia) in 0.5% MMR and injected at the four-cell stage with 10 nl of a solution containing 0.5–1 ng of RNA. Injections were done in the animal pole of one ventral and one dorsal blastomere for analysing Lgl activity and in all four blastomeres for immunoprecipitation analysis. Morpholino antisense oligonucleotides (GeneTools) were injected at 20–40 ng per injection. *XdshMO* has been described²⁶. *LglMO* had the following sequence: 5′-TGCAGCCACAGCTCCCGCACTCACAC-3′; the control MO was of a similar base composition, but had a different sequence: 5′-GTCAGTGTCTGAGGCTCATTCTCTGG-3′. In all experiments, embryos were analysed at stages 10–11. Each experiment was repeated at least three times.

Immunoprecipitation, western blotting and immunofluorescence. Specific antibodies against LglC, LglN and *Dsh* were obtained from rabbits immunized with bacterially produced recombinant fusions of glutathione S-transferase (GST) with the carboxy-terminal amino acids 949–1,035 of *Xenopus* Lgl1, the N-terminal amino acids 2–164 of *Xenopus* Lgl1 and amino acids 79–249 of human *Dvl2*, respectively, and were affinity-purified on corresponding antigens. For immunochemical analysis, injected embryos were lysed in 500 μ l of lysis buffer (1% Triton X-100, 50 mM TrisHCl, 50 mM NaCl, 1 mM EDTA, 0.1 mM phenylmethylsulphonyl fluoride, 10 mM NaF, 1 mM Na₃VO₄; pH 7.5) at early gastrula stages. Immunoprecipitation and western blot analysis were done as described³⁰. For immunoprecipitation, 20 μ l of 9E10 (anti-Myc) hybridoma supernatant, 4 μ l of anti-*Dsh*, or 4 μ l of anti-GST affinity-purified rabbit polyclonal antibodies were used per sample. The equivalent of one-half of an embryo was loaded per lane for embryo lysates, and the equivalent of 15–20 embryos was used for immunoprecipitated proteins. Blots were probed with 9E10 or 12CA5 (anti-HA) hybridoma supernatants, anti-*Dsh* polyclonal and anti-GFP monoclonal antibodies. For cryosectioning, injected embryos were devitellinized at stages 9–10, fixed in Dent's fixative and then 100% methanol, rehydrated in PBS, embedded in a 15% gelatin/15% sucrose solution, frozen on dry ice and sectioned at 12 μ m in a Leica Cryostat. Sections were immunostained as described²⁵. We used the following primary antibodies: anti-aPKC (sc-216, Santa Cruz Biotech; diluted 1/200), anti-occludin (1/100), anti-LglC (1/200) and anti- β 1 integrin (8C8, Developmental Hybridoma Bank; 1/100). Sections were double-stained with a monoclonal anti-GFP antibody (Clontech; 1/200) to identify injected cells, and visualized with fluorescein isothiocyanate (FITC)-conjugated anti-mouse (1/200) or Cy3-conjugated

anti-rabbit (1/100) secondary antibodies (Jackson) on a BioRad confocal microscope. A representative section of an experimental group of 6–9 sectioned embryos is presented.

Fly stocks and clonal analysis. *yw, dsh^{V26}, FRT18A/FM7* virgin females were crossed to *ALZ-FRT18A/y; hsf1p38* males. F₁ females aged 3–5 d lacking FM7 were heat-shocked at 37 °C for 2 h on two consecutive days to induce follicle cell clones. After heat-shock, flies were aged 2–3 d, ovaries were dissected in PBS, fixed in 4% paraformaldehyde for 30 min at room temperature, and co-stained with antibodies against β -galactosidase (β -gal; diluted 1/1,000; Cappel) and against Lgl (diluted 1/100)¹³. Dsh mutant follicle cells were identified by the absence of β -galactosidase.

Received 12 February; accepted 22 July 2005.

- Wodarz, A. Establishing cell polarity in development. *Nature Cell Biol.* **4**, E39–E44 (2002).
- Ohshiro, T., Yagami, T., Zhang, C. & Matsuzaki, F. Role of cortical tumour-suppressor proteins in asymmetric division of *Drosophila* neuroblast. *Nature* **408**, 593–596 (2000).
- Bilder, D., Li, M. & Perrimon, N. Cooperative regulation of cell polarity and growth by *Drosophila* tumour suppressors. *Science* **289**, 113–116 (2000).
- Musch, A. et al. Mammalian homolog of *Drosophila* tumour suppressor lethal (2) giant larvae interacts with basolateral exocytic machinery in Madin–Darby canine kidney cells. *Mol. Biol. Cell* **13**, 158–168 (2002).
- Etienne-Manneville, S. & Hall, A. Cdc42 regulates GSK-3 β and adenomatous polyposis coli to control cell polarity. *Nature* **421**, 753–756 (2003).
- Benton, R. & St Johnston, D. *Drosophila* PAR-1 and 14–3-3 inhibit Bazooka/ PAR-3 to establish complementary cortical domains in polarized cells. *Cell* **115**, 691–704 (2003).
- Chia, W. & Yang, X. Asymmetric division of *Drosophila* neural progenitors. *Curr. Opin. Genet. Dev.* **12**, 459–464 (2002).
- Ohno, S. Intercellular junctions and cellular polarity: the PAR–aPKC complex, a conserved core cassette playing fundamental roles in cell polarity. *Curr. Opin. Cell Biol.* **13**, 641–648 (2001).
- Shi, S. H., Jan, L. Y. & Jan, Y. N. Hippocampal neuronal polarity specified by spatially localized mPar3/mPar6 and PI 3-kinase activity. *Cell* **112**, 63–75 (2003).
- Nakaya, M. et al. Meiotic maturation induces animal-vegetal asymmetric distribution of aPKC and ASIP/PAR-3 in *Xenopus* oocytes. *Development* **127**, 5021–5031 (2000).
- Doe, C. Q. & Bowerman, B. Asymmetric cell division: fly neuroblast meets worm zygote. *Curr. Opin. Cell Biol.* **13**, 68–75 (2001).
- Plant, P. J. et al. A polarity complex of mPar-6 and atypical PKC binds, phosphorylates and regulates mammalian Lgl. *Nature Cell Biol.* **5**, 301–308 (2003).
- Betschinger, J., Mechtler, K. & Knoblich, J. A. The Par complex directs asymmetric cell division by phosphorylating the cytoskeletal protein Lgl. *Nature* **422**, 326–330 (2003).
- Yamanaka, T. et al. Mammalian Lgl forms a protein complex with PAR-6 and aPKC independently of PAR-3 to regulate epithelial cell polarity. *Curr. Biol.* **13**, 734–743 (2003).
- Hutterer, A., Betschinger, J., Petronczki, M. & Knoblich, J. A. Sequential roles of Cdc42, Par-6, aPKC, and Lgl in the establishment of epithelial polarity during *Drosophila* embryogenesis. *Dev. Cell* **6**, 845–854 (2004).
- Klezovitch, O., Fernandez, T. E., Tapscott, S. J. & Vasioukhin, V. Loss of cell polarity causes severe brain dysplasia in Lgl1 knockout mice. *Genes Dev.* **18**, 559–571 (2004).
- Lehman, K., Rossi, G., Adamo, J. E. & Brennwald, P. Yeast homologues of tomosyn and lethal giant larvae function in exocytosis and are associated with the plasma membrane SNARE, Sec9. *J. Cell Biol.* **146**, 125–140 (1999).
- Baas, A. F. et al. Complete polarization of single intestinal epithelial cells upon activation of LKB1 by STRAD. *Cell* **116**, 457–466 (2004).
- Eaton, S. Cell biology of planar polarity transmission in the *Drosophila* wing. *Mech. Dev.* **120**, 1257–1264 (2003).
- Mlodzik, M. Planar cell polarization: do the same mechanisms regulate *Drosophila* tissue polarity and vertebrate gastrulation? *Trends Genet.* **18**, 564–571 (2002).
- Tada, M. & Smith, J. C. Xwnt11 is a target of *Xenopus* Brachyury: regulation of gastrulation movements via Dishevelled, but not through the canonical Wnt pathway. *Development* **127**, 2227–2238 (2000).
- Wallingford, J. B. et al. Dishevelled controls cell polarity during *Xenopus* gastrulation. *Nature* **405**, 81–85 (2000).
- Cadigan, K. M. & Nusse, R. Wnt signalling: a common theme in animal development. *Genes Dev.* **11**, 3286–3305 (1997).
- Chalmers, A. D., Strauss, B. & Papalopulu, N. Oriented cell divisions asymmetrically segregate aPKC and generate cell fate diversity in the early *Xenopus* embryo. *Development* **130**, 2657–2668 (2003).
- Chalmers, A. D. et al. Crumbs3 and Lgl2 control apicobasal polarity in early vertebrate development. *Development* **132**, 977–986 (2005).
- Sheldahl, L. C. et al. Dishevelled activates Ca²⁺ flux, PKC, and CamKII in vertebrate embryos. *J. Cell Biol.* **161**, 769–777 (2003).
- Xu, T. & Rubin, G. M. Analysis of genetic mosaics in developing and adult *Drosophila* tissues. *Development* **117**, 1223–1237 (1993).
- Wu, J., Klein, T. J. & Mlodzik, M. Subcellular localization of frizzled receptors, mediated by their cytoplasmic tails, regulates signalling pathway specificity. *PLoS Biol.* **2**, E158 (2004).
- Bellaiche, Y., Beaudoin-Massiani, O., Stuttem, I. & Schweisguth, F. The planar cell polarity protein Strabismus promotes Pins anterior localization during asymmetric division of sensory organ precursor cells in *Drosophila*. *Development* **131**, 469–478 (2004).
- Brott, B. K. & Sokol, S. Y. A vertebrate homologue of the cell cycle regulator Dbf4 is a Wnt inhibitor required for heart development. *Dev. Cell* **8**, 703–715 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Itoh, J. Green, J. LeBlanc-Straceski, B. Lake and S. Dhawan for discussions and comments on the manuscript; J. Knoblich for anti-Lgl antibody; S. Citi for anti-occludin antibodies; A. Chalmers for the immunohistochemistry protocol; B. Brott for initial work on the yeast two-hybrid screen; and A. Djiane for help with the *Drosophila* studies. This work was supported by NIH grants (to S.Y.S. and M.M.) and an NIH training grant (to G.L.D.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.Y.S. (sergei.sokol@mssm.edu).

A heterodimeric complex that promotes the assembly of mammalian 20S proteasomes

Yuko Hirano¹, Klavs B. Hendil², Hideki Yashiroda¹, Shun-ichiro Iemura³, Ryoichi Nagane³, Yusaku Hioki³, Tohru Natsume³, Keiji Tanaka¹ & Shigeo Murata^{1,4}

The 26S proteasome is a multisubunit protease responsible for regulated proteolysis in eukaryotic cells^{1,2}. It comprises one catalytic 20S proteasome and two axially positioned 19S regulatory complexes³. The 20S proteasome is composed of 28 subunits arranged in a cylindrical particle as four heteroheptameric rings, $\alpha_{1-7}\beta_{1-7}\beta_{1-7}\alpha_{1-7}$ (refs 4, 5), but the mechanism responsible for the assembly of such a complex structure remains elusive. Here we report two chaperones, designated proteasome assembling chaperone-1 (PAC1) and PAC2, that are involved in the maturation of mammalian 20S proteasomes. PAC1 and PAC2 associate as heterodimers with proteasome precursors and are degraded after formation of the 20S proteasome is completed. Overexpression of PAC1 or PAC2 accelerates the formation of precursor proteasomes, whereas knockdown by short interfering RNA impairs it, resulting in poor maturation of 20S proteasomes. Furthermore, the PAC complex provides a scaffold for α -ring formation and keeps the α -rings competent for the subsequent formation of half-proteasomes. Thus, our results identify a mechanism for the correct assembly of 20S proteasomes.

It is presumed that assembly of 20S proteasomes starts by the spontaneous formation of α -rings^{6–8}; however, the exact mechanism responsible for α -ring formation remains elusive. Seven β -subunits, some of which are in precursor forms, are arranged on the α -ring to form a complex named the ‘half-proteasome’, which consists of one α -ring, one β -ring and the chaperone protein Ump1. To complete maturation of the 20S proteasome, two half-proteasomes dimerize, the propeptides of β -subunits are removed and Ump1 is degraded^{9–14}. This model is based mainly on studies in yeast. In mammals, POMP or Proteasembilin, a homologue of yeast Ump1 referred to here as human Ump1 (hUmp1), is also implicated in assembly of 20S proteasomes^{15–17}. However, the biogenesis of 20S proteasomes remains largely elusive, especially in mammalian cells.

To identify proteins that interact with mammalian proteasomes, β 1i subunits with a Flag tag were expressed in cells and anti-Flag immunoprecipitates were analysed by liquid chromatography coupled with tandem mass spectrometry¹⁸. We identified hUmp1 in addition to almost all of the subunits of 20S proteasomes and 19S regulatory complexes. We also identified two molecules with previously unknown relevance to proteasomes. One was Down syndrome critical region 2 (DSCR2), a small leucine-rich protein of 288 amino acids¹⁹ that we have renamed PAC1. The other was a protein of 264 amino acids known as hepatocellular carcinoma associated gene 3 (HCCA3) (ref. 20), which we have renamed PAC2. Both PAC1 and PAC2 are ubiquitously expressed in mammals^{19,20}.

First, we confirmed that these molecules interact physically with proteasomes by transfecting Flag-PAC1 or Flag-PAC2 into

HEK293T cells. The association increased on treatment of the cells with MG132, a proteasome inhibitor, in line with the increase in PAC1 and PAC2 expression (Supplementary Fig. 1a). Next, extracts of HeLa cells stably expressing Flag-PAC1 or Flag-PAC2 were fractionated by 8–32% glycerol gradient centrifugation. Both Flag-PAC1 and Flag-PAC2 were observed mainly in fractions containing sediments of precursor forms of proteasomes, as shown by the co-sedimentation of hUmp1, the unprocessed β 1i (pro- β 1i) subunit and α -subunits, and by the lack of chymotrypsin-like activity. Both Flag-PAC1 and Flag-PAC2 effectively co-precipitated with subunits from fraction 10 (Supplementary Fig. 1b, d). Moreover, even in fraction 16, which contained predominantly mature 20S proteasomes, they precipitated mainly with pro- β 1i (Supplementary Fig. 1d), confirming that PAC1 and PAC2 associate specifically with precursor 20S proteasomes. Notably, the concentrations of precursor proteasomes were increased in both transfected cell lines (Supplementary Fig. 1c).

To examine the behaviour of endogenous PAC1 and PAC2 in detail, extracts from HEK293T cells were separated by lower density (4–24%) glycerol gradient centrifugation to resolve the precursor complexes. PAC1 and PAC2 were distributed mostly in the precursor fractions (Fig. 1a). Notably, the peaks of PAC1 and PAC2 were located in a fraction (fraction 12) lighter than that of the half-proteasomes (fraction 16), which contained hUmp1 and pro- β 2. Moreover, the peaks of α 5– α 7 in precursor fractions were also located in fraction 12. Treatment with MG132 resulted in an accumulation of PAC1 and PAC2 in 20S proteasome fractions (Fig. 1a, right). The association of PAC1 and PAC2 with proteasomes was observed in fractions 12, 16 and 22 (Fig. 1b). When cells were treated with MG132, greater amounts of PAC1 and PAC2 were precipitated from fraction 22. Neither α 4 nor α 6 was associated with pro- β 2 or hUmp1 in fraction 12. These results indicate that PAC1 and PAC2 form a complex with precursor 20S proteasomes before hUmp1 and the pro- β subunits are recruited, and suggest that PAC1 and PAC2 are chaperones for the maturation of 20S proteasomes and are released from or degraded by the newly assembled 20S proteasomes, analogous to the role of Ump1 in yeast¹⁴.

To determine the composition of the peak of α -subunits that contained PAC1 and PAC2, fractions 12 and 16 were immunoprecipitated with antibodies against α 6 and separated by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). Fraction 12 contained all seven α -subunits but no β -subunits, some of which were apparently detected in fraction 16 (Fig. 1c). Immunoblot analysis confirmed that all α -subunits except α 1, which was difficult to distinguish by immunoblotting, were present in fraction 12 (Fig. 1d). The size of this complex (Fig. 1a), coupled with the absence of pro- β subunits or hUmp1 (Fig. 1a–c), means that it is

¹Laboratory of Frontier Science, Core Technology and Research Center, Tokyo Metropolitan Institute of Medical Science, Bunkyo-ku, Tokyo 113-8613, Japan. ²Institute of Molecular Biology and Physiology, University of Copenhagen, 13 Universitetsparken, DK 2100 Copenhagen, Denmark. ³National Institute of Advanced Industrial Science and Technology, Biological Information Research Center, Kohtoh-ku, Tokyo 135-0064, Japan. ⁴PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan.

most probably a ring of all seven α -subunits, namely an α -ring. Immunoprecipitation in lower salt conditions showed that PAC1 and PAC2 are near-stoichiometric components of α -rings and that the association of PAC1 and PAC2 with α -subunits is salt labile (Fig. 1e, f). PAC1 was detected at a wide range of isoelectric point (pI) values, suggesting that it undergoes posttranslational modification (Fig. 1f). These results suggest that the PAC1–PAC2 complex and hUmp1 are distinct entities that work at different points in 20S proteasome assembly, and that PAC1 and PAC2 function as

chaperone-like molecules at an earlier stage of 20S proteasome assembly relative to hUmp1.

Next, we characterized the interaction between PAC1 and PAC2. Coexpression of PAC1 and PAC2 in *Escherichia coli* and *in vitro* cotranscription–translation (IVTT) indicated that the two proteins bind directly (Supplementary Fig. 2a, b). Furthermore, PAC1 tagged with glutathione S-transferase (GST) pulled down PAC2 tagged with haemagglutinin A (HA) but not HA–PAC1, whereas GST–PAC2 pulled down HA–PAC1 but not HA–PAC2 *in vitro* (Supplementary Fig. 2b), indicating that PAC1 and PAC2 form hetero-oligomers but not homo-oligomers. No direct interaction between the PAC complex and hUmp1 was detected (Supplementary Fig. 2c). To determine the stoichiometry of the PAC complex, we coexpressed 3 $\times$ Flag–PAC1 and 6 $\times$ His–PAC2 in *E. coli* and purified the complex. PAC1 and PAC2 formed a complex at 1:1 stoichiometry with a relative molecular mass (M_r) corresponding to bovine serum albumin (67,000; Fig. 2a, b), indicating that the complex is a heterodimer.

To clarify further the nature of the PAC complex, we examined the half-lives of PAC1 and PAC2 by pulse-chase experiments. Both PAC1 and PAC2 turned over rapidly with similar half-lives of about 40 min (Fig. 2c). Treating the cells with MG132 markedly prolonged their half-lives, indicating that the PAC heterodimer is degraded by proteasomes. Because assembly of 20S proteasomes is complete within 1 h (ref. 21), the half-life of the PAC complex is consistent with the complex functioning as a chaperone for proteasome assembly and with its degradation on the completion of 20S proteasomes assembly.

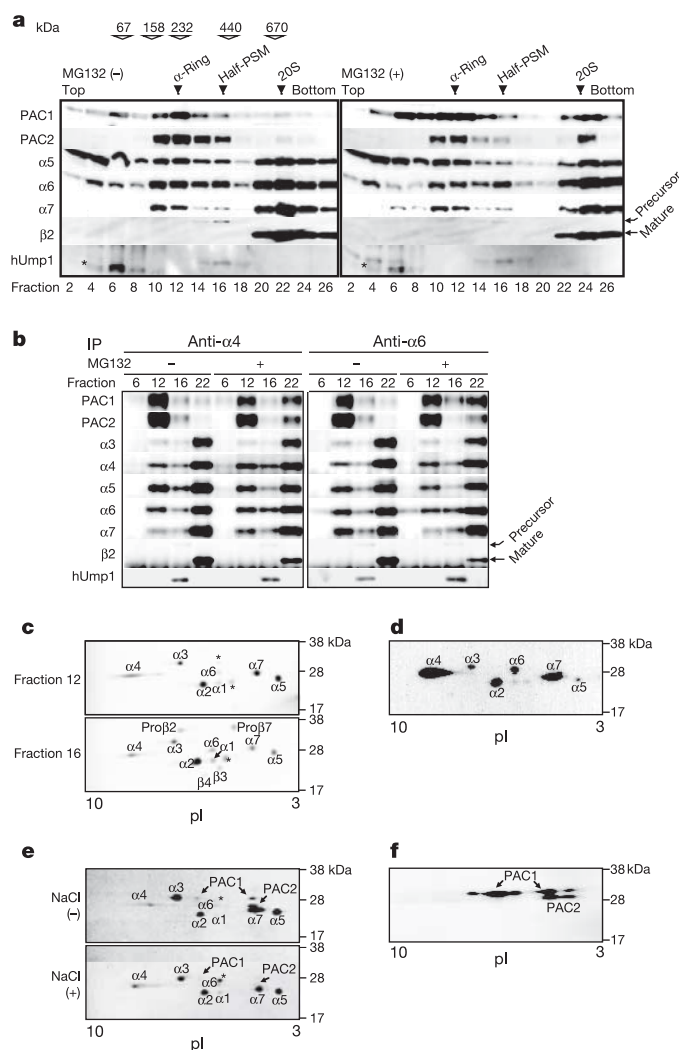


Figure 1 | PAC1 and PAC2 associate with precursor proteasomes.

a, Glycerol gradient centrifugation (4–24%) of HEK293T cell extracts untreated or treated with MG132. Fractions were immunoblotted for the indicated proteins. Size markers and subcomplexes of proteasomes are indicated by open and filled arrowheads, respectively. Half-PSM indicates half-proteasomes. Asterisks indicate nonspecific bands. **b**, Fractions from **a** were immunoprecipitated with antibody against $\alpha 4$ or $\alpha 6$ and then subjected to immunoblotting. **c–f**, Fractions 12 (**c–f**) and 16 (**c**) from **a** were immunoprecipitated with beads conjugated to antibody against $\alpha 6$, washed with buffer A containing 150 mM (**c, d**), 0 mM or 50 mM (**e, f**) NaCl, eluted with glycine-HCl, and resolved by 2D-PAGE with silver (**c**) or Coomassie blue (**e**) staining. Asterisks denote unidentified spots. The top gel in **c** was immunoblotted with MCP231 and MCP34 antibodies against α -subunits and $\alpha 4$, respectively (**d**). The top gel in **e** was immunoblotted with antibodies against PAC1 and PAC2 (**f**). The non-uniformity of the spot intensity (**c, e**) may be due to a staining artefact because even in the half-proteasome fraction (fraction 16), which should contain all α -subunits in equal amounts, the spot intensities of α -subunits varied, resembling the pattern of fraction 12.

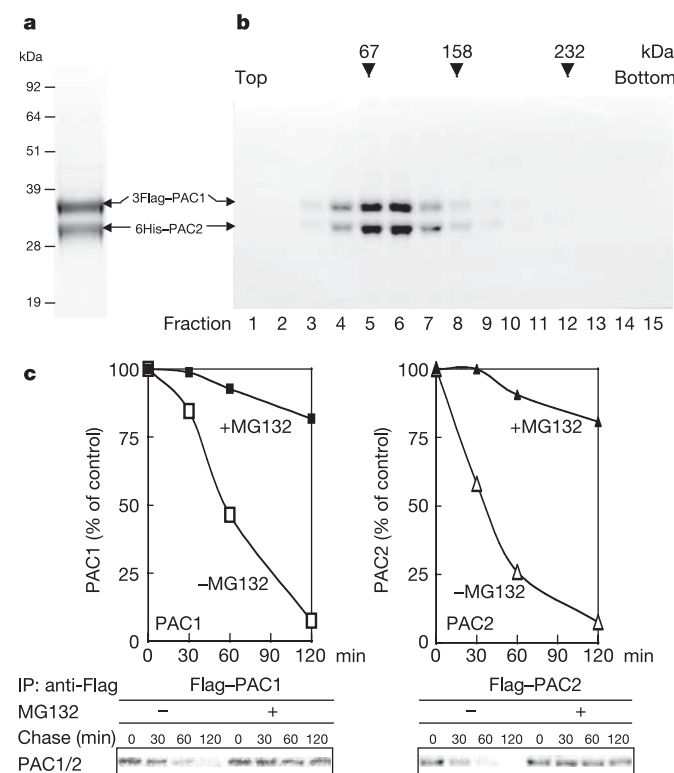


Figure 2 | The PAC1–PAC2 heterodimer is rapidly degraded by proteasomes.

a, Coomassie blue staining of a copurified complex of 3 $\times$ Flag–PAC1 and 6 $\times$ His–PAC2 expressed in bacterial cells. **b**, The purified PAC complex in **a** was separated by 4–17% glycerol gradient centrifugation and subjected to SDS–PAGE with Coomassie staining. Arrowheads indicate size markers. **c**, Half-lives of PAC1 and PAC2. HeLa cells stably transfected with Flag–PAC1 or Flag–PAC2 were radiolabelled and chased in the presence or absence of MG132. Bottom panels show autoradiography; top panels show quantitative analysis of the bands.

To clarify the role of PAC1 and PAC2 in the assembly of the 20S proteasome *in vivo*, we used short interfering RNA (siRNA) to knock down the expression of PAC1 and PAC2. Knockdown of PAC1 resulted in loss of both PAC1 and PAC2 protein. Knockdown of PAC2 was also associated with a decrease in PAC1 protein (Fig. 3a), indicating that PAC1 and PAC2 are stable only when they form a heterodimer. Both PAC1- and PAC2-knockdown cells showed reduced proteolytic activity, as indicated by an assay of the anti-enzyme-dependent degradation of ornithine decarboxylase (Supplementary Fig. 3a). Consequently, PAC-knockdown cells accumulated polyubiquitin-conjugated proteins, were sensitive to stress such as Cd^{2+} , and showed slow growth (Supplementary Fig. 3b–d).

We subjected the PAC-knockdown cells, as well as control and hUmp1-knockdown cells, to 4–24% glycerol gradient analysis. Notably, α -rings were hardly detected in either the PAC1- or the PAC2-knockdown cells (Fig. 3b). Instead, the α -subunits accumulated in fractions corresponding to half-proteasomes. This accumulation was not accompanied by an increase in pro- β 2, pro- β 5 or hUmp1, however, suggesting that the half-proteasomes were not normal. To confirm this notion, fraction 16 from the knockdown cells was immunoprecipitated with antibody against α 6. Even though nearly equal amounts of α -subunits were loaded in the different samples, pro- β 2, pro- β 5 and hUmp1 were detected in much smaller amounts in PAC-knockdown cells (Fig. 3c), indicating that fraction 16 in PAC-knockdown cells contained mostly abnormally assembled α -subunits. This abnormal complex did not contain Rpt subunits, the components of 19S regulatory particles (Supplementary Fig. 3e, f), precluding the possibility that the mobility shift of α -subunits in PAC-knockdown cells was due to the premature association of α -subunits with Rpt subunits. On the basis of their sizes, these complexes are probably dimers of α -rings.

In hUmp1-knockdown cells, by contrast, we observed a marked reduction in 20S proteasomes but apparently normal α -rings and half-proteasomes, demonstrating the crucial role of hUmp1 in the dimerization of half-proteasomes. There was a strong increase in the free forms of some α -subunits in hUmp1-knockdown cells and a

moderate increase in PAC-knockdown cells (Supplementary Fig. 3g). Assays of peptidase activities showed a significant reduction in activity of both the 20S and the 26S proteasome fractions in PAC-knockdown cells, although the effect of hUmp1 knockdown was more intense (Supplementary Fig. 3h). These data show definitively that the PAC complex has a pivotal role in the assembly of 20S proteasomes, specifically in keeping α -rings competent for the subsequent formation of half-proteasomes.

To elucidate the mechanism of PAC complex function, we tested the direct association of the complex with all of the subunits of 20S proteasomes. The PAC complex specifically interacted with α 5 and α 7, but not with other α -subunits or with any of the β -subunits *in vitro* (Supplementary Fig. 4a). Because Fig. 1a shows that the PAC1–PAC2 complex is found not only in α -ring fractions but also in lighter fractions, we considered whether it is involved in α -ring assembly. Immunoprecipitation with an antibody against Flag after the coexpression of all seven α -subunits, of which α 5 was Flag-tagged, by IVTT showed that all α -subunits co-precipitated with Flag- α 5 in larger amounts in the presence of the PAC complex than in its absence (Supplementary Fig. 4b). Immunoprecipitation with anti-Flag antibody after the coexpression of Flag-PAC1, PAC2 and α -subunits showed that PAC1 precipitated not only α 5 and α 7 but also all of the other α -subunits (Supplementary Fig. 4c), implying that it has a role in attracting α -subunits to each other.

We examined these interactions under more physiological conditions. Extracts of 293T cells that stably express Flag-PAC1 were separated by 4–24% glycerol gradient, and fractions corresponding to early α -subunit assembly intermediates and α -rings (fractions 8 and 12 in Fig. 1a, respectively) were immunoprecipitated with anti-Flag antibody and subjected to 2D-PAGE. PAC1 in fraction 12 co-precipitated all seven α -subunits, whereas PAC1 in fraction 8 co-precipitated several unidentified spots other than α -subunits, which made it difficult to identify α -subunits except for α 5 and α 7 by Coomassie staining (Fig. 4a, left). Immunoblot analysis showed that all α -subunits were present in the α -ring fraction, although α 1 was difficult to distinguish (Fig. 4a, right), consistent with the findings in Fig. 1c. In contrast, PAC1 in fraction 8 co-precipitated a restricted set

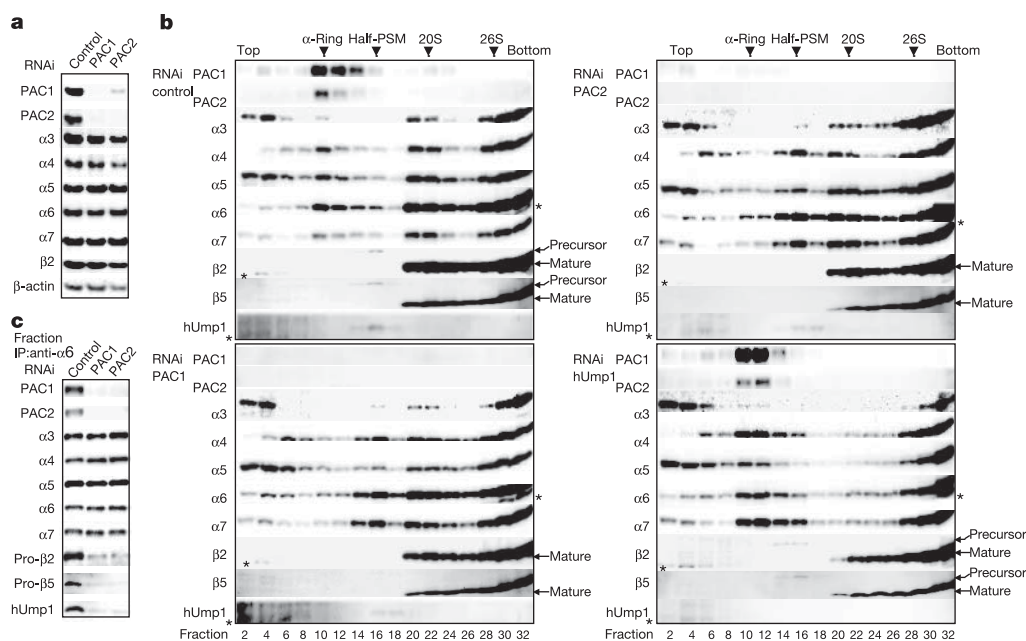


Figure 3 | siRNA-mediated knockdown of PAC1 and PAC2 impairs proteasome assembly. **a–c**, siRNA targeting PAC1 or PAC2, or control siRNA, was transfected into HEK293T cells. Knockdown of hUmp1 was also analysed in **b**. Whole-cell extracts (**a**), fractions separated by 4–24% glycerol

gradient centrifugation (**b**), and immunoprecipitates obtained from fraction 16 in **b** with antibodies against α 6 (**c**) were immunoblotted for the indicated proteins. Asterisks indicate nonspecific bands.

of α -subunits in which $\alpha 3$ and $\alpha 4$ were hardly detected. These results indicate that there is a hierarchy among α -subunits in their incorporation into α -rings, and that the PAC complex associates with the α -subunits before α -rings are complete, and functions as a scaffold for α -ring assembly.

Finally, we tested whether the complex of α -subunits in fraction 12 is a unique species. The affinity-purified complex from fraction 12 was subjected to native-PAGE. We found that the complex had a unique electrophoretic mobility (Fig. 4b). Moreover, the complex was eluted with a single sharp peak by anion-exchange chromatography (Fig. 4c). Thus, this complex is a unique species

biochemically and is a genuine α -ring rather than a group of heterogeneous and incomplete α -ring precursors.

Our present work provides a model in which the chaperone complex PAC1–PAC2 mediates the formation of α -rings, keeps the rings competent for half-proteasome formation, and is required for proper proteasome maturation and cellular integrity (Fig. 4d). (See Supplementary Discussion for a more detailed description.)

METHODS

See Supplementary Methods for procedures used in the experiments in Supplementary Figs 1–4.

DNA constructs and cell culture. We synthesized cDNAs encoding PAC1, PAC2, hUmp1 and proteasome α - and β -subunits from total RNA isolated from HeLa cells using Superscript II (Invitrogen). PCR was carried out on the cDNA with Pyrobest DNA polymerase (Takara). All of the amplified fragments were cloned into pCDNA3.1 (Invitrogen) and sequenced for confirmation. For expression of GST fusion proteins, the cDNAs were subcloned into pGEX6P-1 (Amersham). Transfections of 293T cells were done with Fugene 6 (Roche). Stable transfections of HeLa cells or 293T cells were done with Lipofectamine 2000 (Invitrogen), and the cells were selected with 1 mg ml^{-1} of G418 or $5 \mu\text{g ml}^{-1}$ of puromycin, respectively. We used $20 \mu\text{M}$ MG132 (Peptide Institute) to inhibit proteasome activities 2 h before the cells were collected.

Protein extracts, immunological analysis and antibodies. Cells were lysed in ice-cold buffer A containing 50 mM Tris-HCl (pH 7.5), 0.5% (v/v) Nonidet P40, 1 mM dithiothreitol (DTT) and 2 mM ATP, and the extracts were clarified by centrifugation at $20,000g$ for 10 min at 4°C . SDS-PAGE (12% gel or 4–12% gradient Bis-Tris gel; Invitrogen) and native-PAGE (3–8% gradient Tris-acetate gel; Invitrogen) were done in accordance with the manufacturer's instructions. The separated proteins were transferred onto polyvinylidene difluoride membrane and reacted with the indicated antibody. Development was done with Western Lighting reagent (Perkin Elmer). Polyclonal antibodies against hUmp1, PAC1 and PAC2 were raised in rabbits using a synthetic peptide ($\text{E}_{115}\text{DILNDPSQSE}_{125}$), and recombinant PAC1 and PAC2 protein, respectively. PAC1 and PAC2 were produced and purified as GST fusion proteins, and GST was removed by PreScission protease (Amersham).

Antibodies against proteasome $\alpha 3$ subunit (MCP257), $\alpha 4$ (MCP34), $\alpha 5$ (MCP196), $\alpha 6$ (MCP20), $\alpha 7$ (MCP72), $\beta 2$ (MCP168) and α -subunits (MCP231, which reacts with all α -subunits except $\alpha 4$) were purchased from BioMol. Antibodies against $\beta 5$ (P93250), $\beta 6$ (P93199) and $\beta 1$ were prepared as described²². We used antibodies against the Flag tag (Sigma) and β -actin (Chemicon) and horseradish peroxidase (HRP)-conjugated rabbit anti-mouse and goat anti-rabbit IgG (Jackson ImmunoResearch) for immunodetection. For immunoprecipitation of the Flag epitope, we used M2 agarose (Sigma). For immunoprecipitation of proteasomes, we used antibody MCP34 or MCP20 bound to protein G Sepharose (Amersham). In the experiments in Fig. 1c–f, we used MCP20 crosslinked to NHS-activated Sepharose (Amersham). These beads were added to the extracts, mixed under constant rotation for 2 h at 4°C , washed four times with buffer A (except in the experiments in Fig. 1c–f), and boiled in SDS sample buffer, or eluted with $100 \mu\text{g ml}^{-1}$ of Flag peptides (Sigma) or with 0.2 M glycine-HCl (pH 2.8). Densitometric analysis was done with Image Gauge software (Fujifilm). 2D-PAGE was done as described²³.

Glycerol gradient analysis. Samples and molecular weight markers (Amersham) were fractionated by 4–17% (v/v), 4–24% (v/v) or 8–32% (v/v) linear glycerol density gradient centrifugation (22 h, $100,000g$) as described²².

Purification of PAC1–PAC2 complex. We coexpressed $3\times \text{FLAG-PAC1}$ and $6\times \text{His-PAC2}$ in *E. coli* using a pRSFDuet-1 vector (Novagen). The cell pellets were lysed in buffer B containing 20 mM sodium phosphate (pH 7.8), 500 mM NaCl and 1.0% Triton X-100, and sonicated. Ni-NTA Sepharose (Qiagen) was added to the extracts, which were then washed with buffer C containing 20 mM sodium phosphate (pH 6.0) and 500 mM NaCl, and eluted with buffer C plus 100 mM imidazole. The eluted products were further purified with M2 agarose and eluted with $100 \mu\text{g ml}^{-1}$ of Flag peptide (Sigma).

Pulse-chase experiments. Cells were incubated with methionine-free medium for 1 h, metabolically labelled with ^{35}S -methionine for 1 h, and then washed and chased for the indicated time. The cell lysates were immunoprecipitated with M2 agarose, fractionated by SDS-PAGE and visualized by autoradiography.

RNAi experiments. siRNAs targeting human PAC1, PAC2 and hUmp1 with the following 19-nucleotide sequences were designed by B-Bridge and synthesized by Dharmacon: PAC1, 5'-CCAGAAGCUUGAAGGGUUU-3'; PAC2, 5'-GCAUAAUUGCUGAAGUGUA-3'; hUmp1, a mixture of 5'-GCAAGUGG ACCUUUUGAAA-3' and 5'-CCUGAGAAUUCUGCUCAA-3'. Control siRNA (Non-specific Control Duplex VIII) was purchased from B-Bridge. Transfections of siRNAs into HEK293T cells were done with Lipofectamine

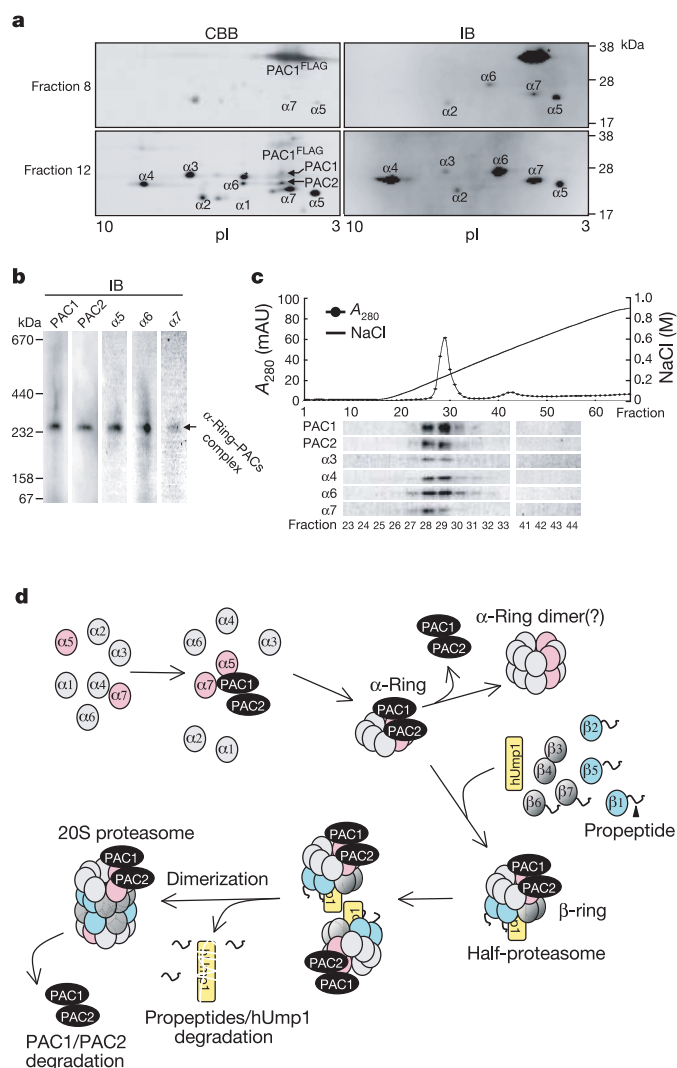


Figure 4 | PAC1–PAC2 provides a scaffold for α -ring formation. **a**, Extracts of HEK293T cells transfected with Flag–PAC1 were fractionated as in Fig. 1a. The Flag–PAC1 complexes from fractions 8 and 12 were purified with M2 agarose, resolved by 2D-PAGE and detected by Coomassie blue staining (left) or immunoblotting as in Fig. 1d (right). Asterisks denote unidentified spots. **b**, **c**, Purified Flag–PAC1 complex from fraction 12 was subjected to native-PAGE (**b**) or anion-exchange chromatography (**c**), followed by immunoblotting. **d**, Multistep model of the ordered assembly of mammalian 20S proteasomes. Some of the newly synthesized free α -subunits bind to the PAC1–PAC2 heterodimer, which provides a scaffold for α -ring formation, thereby suppressing the off-pathway aggregation of α -subunits and keeping α -rings competent for half-proteasome formation. Two half-proteasomes then dimerize, the β -subunits are processed and hUmp1 is degraded. The PAC1–PAC2 complex is subsequently degraded by the newly formed active 20S proteasomes.

2000 at a final concentration of 50 nM in six-well dishes. The cells were analysed 72 h after transfection.

Assay of proteasome activity. Peptidase activity was measured by using a fluorescent peptide substrate, succinyl-Leu-Val-Tyr-7-amido-4-methylcoumarin (Suc-LLVY-MCA), as described²³.

Chromatography. Anion-exchange chromatography was done with a Resource Q column (Amersham). Bound proteins were eluted with a salt gradient of 0–1 M NaCl in a buffer containing 50 mM Tris-HCl (pH 8.0), 1 mM DTT and 5% glycerol.

Received 27 June; accepted 2 August 2005.

- Pickart, C. M. Mechanisms underlying ubiquitination. *Annu. Rev. Biochem.* **70**, 503–533 (2001).
- Glickman, M. H. & Ciechanover, A. The ubiquitin–proteasome proteolytic pathway: destruction for the sake of construction. *Physiol. Rev.* **82**, 373–428 (2002).
- Baumeister, W., Walz, J., Zuhl, F. & Seemuller, E. The proteasome: paradigm of a self-compartmentalizing protease. *Cell* **92**, 367–380 (1998).
- Groll, M. *et al.* Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* **386**, 463–471 (1997).
- Unno, M. *et al.* The structure of the mammalian 20S proteasome at 2.75 Å resolution. *Structure (Camb.)* **10**, 609–618 (2002).
- Zwickl, P., Kleinz, J. & Baumeister, W. Critical elements in proteasome assembly. *Nat. Struct. Biol.* **1**, 765–770 (1994).
- Gerards, W. L. *et al.* The human α -type proteasomal subunit HsC8 forms a double ringlike structure, but does not assemble into proteasome-like particles with the β -type subunits HsDelta or HsBPROS26. *J. Biol. Chem.* **272**, 10080–10086 (1997).
- Yao, Y. *et al.* α 5 subunit in *Trypanosoma brucei* proteasome can self-assemble to form a cylinder of four stacked heptamer rings. *Biochem. J.* **344**, 349–358 (1999).
- Yang, Y., Fruh, K., Ahn, K. & Peterson, P. A. *In vivo* assembly of the proteasomal complexes, implications for antigen processing. *J. Biol. Chem.* **270**, 27687–27694 (1995).
- Chen, P. & Hochstrasser, M. Autocatalytic subunit processing couples active site formation in the 20S proteasome to completion of assembly. *Cell* **86**, 961–972 (1996).
- Schmidtke, G. *et al.* Analysis of mammalian 20S proteasome biogenesis: the maturation of β -subunits is an ordered two-step mechanism involving autocatalysis. *EMBO J.* **15**, 6887–6898 (1996).
- Nandi, D., Woodward, E., Ginsburg, D. B. & Monaco, J. J. Intermediates in the formation of mouse 20S proteasomes: implications for the assembly of precursor β subunits. *EMBO J.* **16**, 5363–5375 (1997).
- Schmidtke, G., Schmidt, M. & Kloetzel, P. M. Maturation of mammalian 20S proteasome: purification and characterization of 13S and 16S proteasome precursor complexes. *J. Mol. Biol.* **268**, 95–106 (1997).
- Ramos, P. C., Hockendorff, J., Johnson, E. S., Varshavsky, A. & Dohmen, R. J. Ump1p is required for proper maturation of the 20S proteasome and becomes its substrate upon completion of the assembly. *Cell* **92**, 489–499 (1998).
- Griffin, T. A., Slack, J. P., McCluskey, T. S., Monaco, J. J. & Colbert, R. A. Identification of proteasemlin, a mammalian homologue of the yeast protein, Ump1p, that is required for normal proteasome assembly. *Mol. Cell. Biol. Res. Commun.* **3**, 212–217 (2000).
- Witt, E. *et al.* Characterisation of the newly identified human Ump1 homologue POMP and analysis of LMP7(β 5i) incorporation into 20S proteasomes. *J. Mol. Biol.* **301**, 1–9 (2000).
- Burri, L. *et al.* Identification and characterization of a mammalian protein interacting with 20S proteasome precursors. *Proc. Natl Acad. Sci. USA* **97**, 10348–10353 (2000).
- Natsume, T. *et al.* A direct nanoflow liquid chromatography-tandem mass spectrometry system for interaction proteomics. *Anal. Chem.* **74**, 4725–4733 (2002).
- Vidal-Taboada, J. M. *et al.* Down syndrome critical region gene 2: expression during mouse development and in human cell lines indicates a function related to cell proliferation. *Biochem. Biophys. Res. Commun.* **272**, 156–163 (2000).
- Bahar, R. *et al.* Growth retardation, polyploidy, and multinucleation induced by Clast3, a novel cell cycle-regulated protein. *J. Biol. Chem.* **277**, 40012–40019 (2002).
- Ahn, K. *et al.* *In vivo* characterization of the proteasome regulator PA28. *J. Biol. Chem.* **271**, 18237–18242 (1996).
- Tanahashi, N. *et al.* Hybrid proteasomes. Induction by interferon- γ and contribution to ATP-dependent proteolysis. *J. Biol. Chem.* **275**, 14336–14445 (2000).
- Murata, S. *et al.* Immunoproteasome assembly and antigen presentation in mice lacking both PA28 α and PA28 β . *EMBO J.* **20**, 5898–5907 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Murakami for the ornithine decarboxylase degradation assay system, K. Furuyama for technical support, and D. Finley for comments on the manuscript. This work was supported by grants from the Japanese Science and Technology Agency (to S.M.), the Ministry of Education, Science and Culture of Japan (to S.M. and K.T.) and the New Energy and Industrial Technology Development Organization (to T.N.). Y.H. was supported by the Japanese Society for the Promotion of Science.

Author Information The sequences for human PAC1 and PAC2 have been deposited in GenBank under accession numbers BR000236 and BR000237, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.M. (smurata@rinshoken.or.jp) or K.T. (tanakak@rinshoken.or.jp).

LETTERS

The histone H3.3 chaperone HIRA is essential for chromatin assembly in the male pronucleus

Benjamin Loppin¹, Emilie Bonnefoy¹, Caroline Anselme², Anne Laurençon¹, Timothy L. Karr³ & Pierre Couble¹

In sexually reproducing animals, a crucial step in zygote formation is the decondensation of the fertilizing sperm nucleus into a DNA replication-competent male pronucleus. Genome-wide nucleosome assembly on paternal DNA implies the replacement of sperm chromosomal proteins, such as protamines, by maternally provided histones^{1,2}. This fundamental process is specifically impaired in *sésame* (*ssm*), a unique *Drosophila* maternal effect mutant that prevents male pronucleus formation³. Here we show that *ssm* is a point mutation in the *Hira* gene, thus demonstrating that the histone chaperone protein HIRA is required for nucleosome assembly during sperm nucleus decondensation. In vertebrates, HIRA has recently been shown to be critical for a nucleosome assembly pathway independent of DNA synthesis that specifically involves the H3.3 histone variant^{4,5}. We also show that nucleosomes containing H3.3, and not H3, are specifically assembled in paternal *Drosophila* chromatin before the first round of DNA replication. The exclusive marking of paternal chromosomes with H3.3 represents a primary epigenetic distinction between parental genomes in the zygote, and underlines an important consequence of the critical and highly specialized function of HIRA at fertilization.

With the exception of its four centromeric regions, the needle-like, extremely condensed *Drosophila* sperm nucleus is devoid of the core histones (H2A, H2B, H3 and H4) that form the nucleosome particle^{6,7}. Indeed, in elongating spermatids, histones are replaced by sperm-specific chromosomal proteins including protamines⁷. At fertilization, *de novo* nucleosome assembly in the male pronucleus is a crucial process that determines the ability of the paternal genome to participate in the formation of the diploid zygote. However, the molecular basis of this process in *Drosophila* is largely unknown⁸. We have previously characterized *sésame* (*ssm*), a unique maternal effect, embryonic-lethal mutation that specifically affects the formation of the male pronucleus³. In eggs laid by homozygous *ssm* females, maternal histones are not deposited in the male pronucleus, which remains abnormally condensed⁶. Consequently, haploid embryos develop with only maternal chromosomes and invariably die before hatching³. This dramatic phenotype suggests that *ssm* is involved in the replacement of protamines with histones in the male pronucleus⁶. To identify the *ssm* gene, we aligned the previously defined *ssm* genetic region³ (7B6–7C1, X chromosome) relative to the annotated genome (Fig. 1a). Of the ten candidate genes predicted in this region, we first tested the *Hira* gene^{9,10} in rescue experiments (Fig. 1b). A single copy of a *Hira* transgene completely rescued the fertility of homozygous *ssm* females, thereby demonstrating that *Hira* is the gene affected by the *ssm* mutation (Supplementary Table 1).

Analysis of *Hira* transcripts by polymerase chain reaction with reverse transcription (RT-PCR) revealed no significant difference in expression between wild-type and *ssm* flies (Supplementary Fig. 1). HIRA proteins are characterized by seven conserved WD repeats

in their amino-terminus, which are predicted to assemble into a β -propeller structure known to mediate protein–protein interactions¹¹. Sequencing of the *Hira* gene amplified from *ssm* flies revealed a single point mutation that results in substitution of the highly conserved arginine 225 with a lysine residue (R225K). This residue is located between the fourth and fifth WD repeats of HIRA (Fig. 1c). The 100% penetrant phenotype of *ssm* arises from this subtle mutation, indicating that this residue sits in a crucial functional domain of the protein.

Recently, HIRA has been shown to be critical for a replication-independent nucleosome assembly pathway, distinct from the major replication-coupled chromatin assembly that occurs during genome duplication⁴. In the decondensing sperm nucleus, chromatin

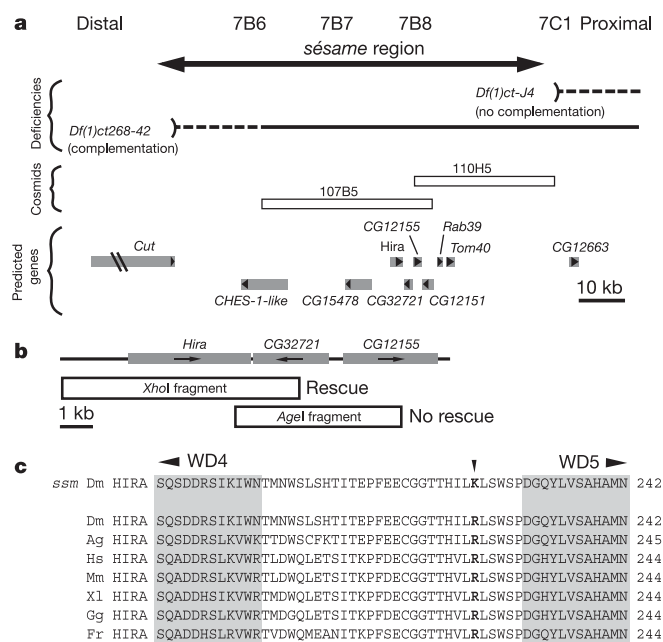


Figure 1 | *Hira* is the *ssm* gene. **a**, Positions of the proximal breakpoints of the deficiencies at the *ssm* locus. Black lines represent the undeleted chromosome regions (uncertainty in breakpoint positions is shown with a broken line). Open boxes, genomic DNA cosmids. Predicted genes²⁹ (grey boxes) are shown with black arrowheads indicating their orientation. **b**, Genomic DNA fragments used in the rescue experiments, shown with the corresponding predicted genes. **c**, Alignments of HIRA proteins from various species showing the R225K substitution in *ssm* (arrowhead, bold). Extremities of WD repeats 4 and 5 are shaded with grey. Dm, *Drosophila melanogaster*; Ag, *Anopheles gambiae*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Xl, *Xenopus laevis*; Gg, *Gallus gallus*; Fr, *Fugu rubripes*.

¹Centre de Génétique Moléculaire et Cellulaire, U.M.R. 5534, C.N.R.S. – Université Claude Bernard, Lyon-1, 69622 Villeurbanne, France. ²U.M.R. INRA/INSA BF21, 69621 Villeurbanne, France. ³Department of Biology and Biochemistry, University of Bath, 4 South Building, Claverton down, Bath BA2 7AY, UK.

assembly takes place before the onset of the first zygotic S phase¹². We thus reasoned that this process could represent a peculiar case of replication-independent nucleosome assembly at the scale of a whole genome. To test this hypothesis, we first studied the distribution of maternal HIRA in fertilized eggs. We established transgenic flies expressing HIRA fused to the 3× Flag tag (HIRA–Flag). Anti-Flag western blot analysis of early embryos from transgenic females revealed a single band corresponding to the expected size of HIRA–Flag (Supplementary Fig. 2). This fusion protein is functional, as the *Hira*–Flag transgene fully rescues the fertility of *ssm* females (Supplementary Table 1).

Immunofluorescence of eggs from transgenic females crossed with wild-type yw^{67c} males revealed bright HIRA–Flag staining in the male nucleus from the initiation of decondensation, when the sperm nucleus is still elongated, until the end of pronuclear formation (Fig. 2a). In contrast, HIRA–Flag was never detected in maternal chromosomes at any stage of meiosis or pronuclear formation (Fig. 2a), and was also absent from embryonic nuclei during early development (not shown). We also verified that sperm from transgenic males did not contribute any detectable level of HIRA–Flag when crossed with wild-type females (not shown). Thus,

the highly specific distribution of maternal HIRA in the male nucleus is consistent with the observed mutant phenotype.

We confirmed this result using antibodies directed against two peptides from the *Drosophila* HIRA protein. In wild-type w^{1118} flies, anti-HIRA antibodies stained the male nucleus exactly like HIRA–Flag did (Fig. 2b). In addition, anti-HIRA antibodies stained the sperm flagellum non-specifically, as is often the case for rabbit antibodies (B.L., unpublished observation). In conclusion, our data show that maternal HIRA very efficiently targets the male nucleus immediately after sperm entry. This highly specific distribution of HIRA on paternal chromatin well before the onset of the first S phase makes a strong case for its potential role in replication-independent nucleosome assembly during male pronucleus formation.

In eggs from mutant females stained with anti-HIRA antibodies, we found that the mutant HIRA was still detected in the male nucleus at all stages of zygote formation (Fig. 2c). We also constructed a version of the *Hira*–Flag transgene containing the R225K substitution (*Hira*^{*ssm*}–Flag). As expected, this construct was unable to rescue the *ssm* phenotype (Supplementary Table 1), despite the fact that HIRA^{*ssm*}–Flag brightly stained the male nucleus in mutant eggs

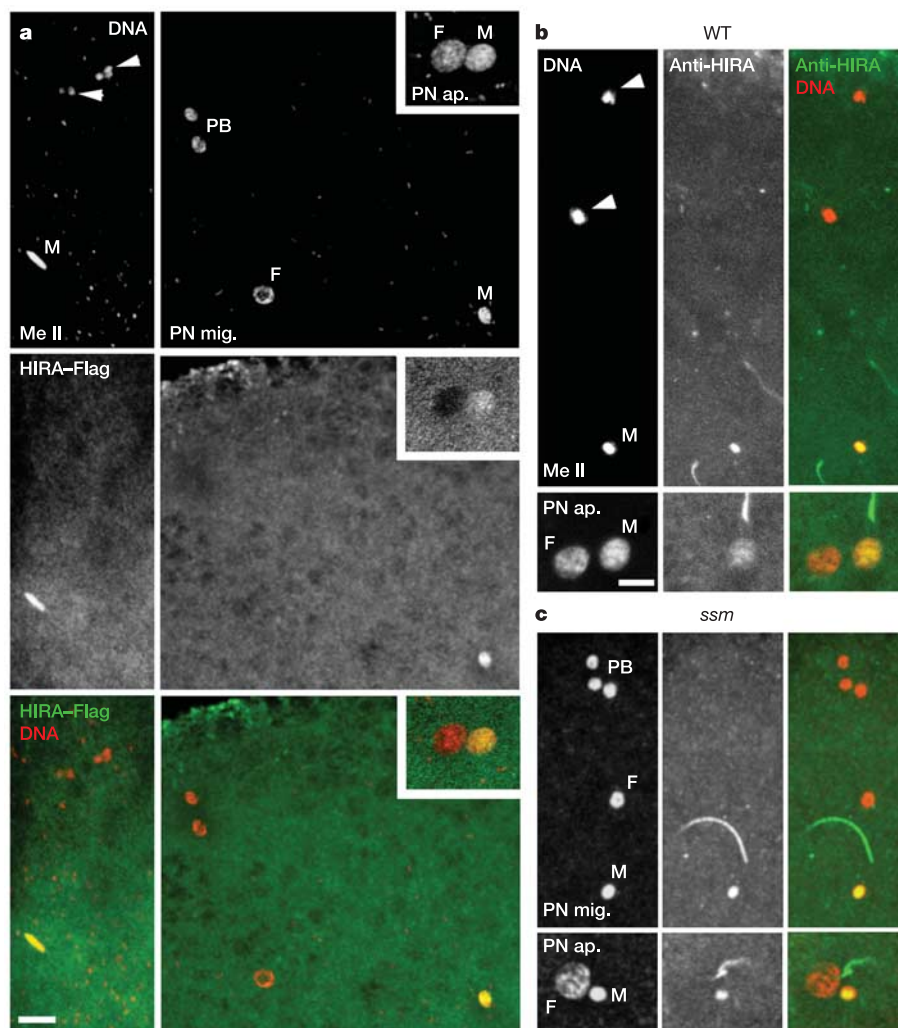


Figure 2 | Maternal HIRA accumulates in the male nucleus. **a**, Confocal images of eggs from *Hira*–Flag transgenic females, staining for DNA (top), HIRA–Flag (middle) or both (bottom). At metaphase of meiosis II (Me II), HIRA–Flag stains the male nucleus (M) but not the maternal chromosomes (arrowheads). During migration of the female pronucleus (PN mig.) and at pronuclear apposition (PN ap.), HIRA–Flag specifically stains the male

nucleus. (Note that only two of the three polar bodies (PB) are visible in the PN mig. section.) **b**, **c**, Anti-HIRA antibodies stain the male nucleus in wild-type (**b**) and *ssm* (**c**) eggs but do not stain maternal chromosomes (arrowheads), polar bodies or the female pronucleus. The sperm flagellum is visible in the green channel. F, female nucleus. Scale bars, 10 μ m.

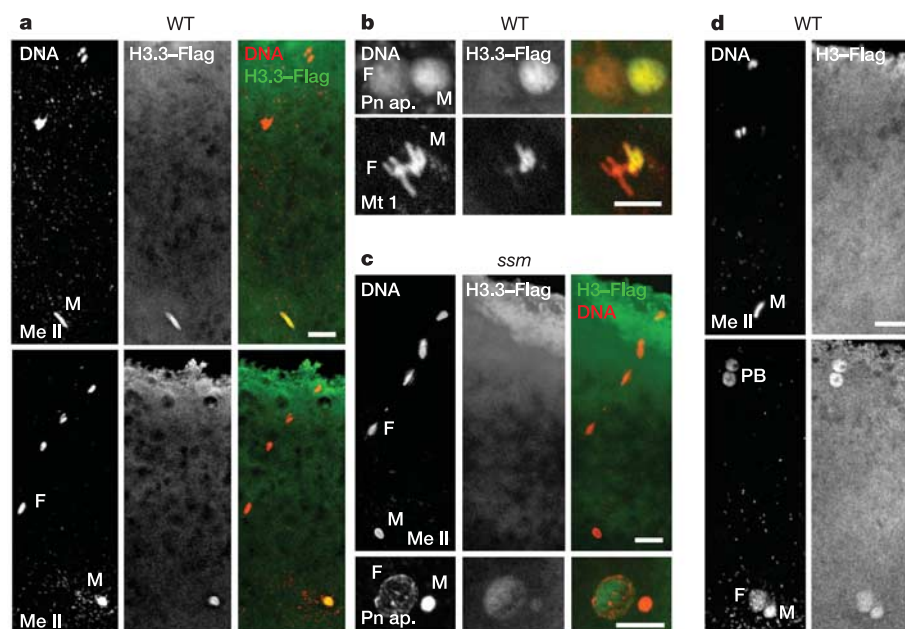


Figure 3 | H3.3 is specifically deposited in paternal chromatin. **a, b,** In wild-type eggs (WT) H3.3-Flag is detected in paternal chromatin from meiosis II (Me II) (top left, metaphase; bottom left, telophase) (**a**) until pronuclear apposition (PN ap.) and first zygotic mitosis (Mt 1) (**b**). **c,** In *ssm* eggs, H3.3-Flag is not detected in the male nucleus (M) during meiosis and is very

weak after pronuclear apposition. **d,** H3-Flag is not detected in the male nucleus during meiosis. After pronuclear apposition, H3-Flag stains both pronuclei and polar bodies (PB; two polar bodies are visible here). F, female nucleus. Scale bars, 10 μ m.

(not shown). In conclusion, the *ssm* mutation does not affect the ability of HIRA to localize to the male nucleus.

Increasing evidence supports an important role for histone H3 variants in specifying modes of nucleosome assembly^{13,14}. Whereas the major H3 histones (H3.1 and H3.2) are synthesized during S phase and are deposited on DNA strictly during DNA replication, the

histone H3 replacement variant H3.3 is synthesized and deposited onto DNA throughout the cell cycle¹⁴. Additionally, H3.3 deposition has been shown to mark transcriptionally active chromatin *in vivo*^{14–16}, suggesting that H3.3 nucleosomes may confer epigenetic inheritance of active chromatin states¹⁷. Recently, the purification from human cells of complexes for replication-coupled deposition of H3.1 and replication-independent deposition of H3.3 has demonstrated that these alternative nucleosome assembly pathways depend on different chromatin assembly factors: the CAF-1 complex during DNA replication or DNA repair, and the HIRA complex, which is independent of DNA synthesis⁵.

The evidence that vertebrate HIRA is specialized for the deposition of H3.3 nucleosomes on DNA prompted us to test the possible implication of this histone variant in male pronucleus formation. H3, the unique *Drosophila* S phase variant, and H3.3 differ by only four amino acids¹⁴, and no specific H3.3 antibody is currently available. We thus established transgenic lines expressing 3X Flag-tagged versions of both histone variants. Both constructs use the regulatory sequences of the *Drosophila* *His3.3A* gene¹⁸, resulting in similar levels of both recombinant histones in eggs (Supplementary Fig. 2).

We first looked at the distribution of H3.3-Flag during fertilization and zygote formation by crossing transgenic females with γw^{67c} males. In all cases, maternal H3.3-Flag specifically accumulated in the male nucleus at all stages of pronuclear formation (Fig. 3a, b). In contrast, the polar bodies and female pronucleus did not stain for H3.3-Flag, except for a weak staining that appeared by the time the pronuclei apposed (Fig. 3b). During the first zygotic mitosis, H3.3-Flag still strongly labelled paternal chromosomes (Fig. 3b), but the staining faded away after a few nuclear divisions (not shown). As expected, H3.3-Flag was absent or barely detected in the male nucleus in eggs from *ssm* mutant females (Fig. 3c).

In clear contrast to H3.3-Flag, maternal H3-Flag was never detected in the decondensing male nucleus (Fig. 3d). However, shortly before pronuclear apposition, both pronuclei and the polar bodies were intensely labelled with H3-Flag (Fig. 3d). This stage corresponds to the onset of the first S phase in all nuclei, as confirmed

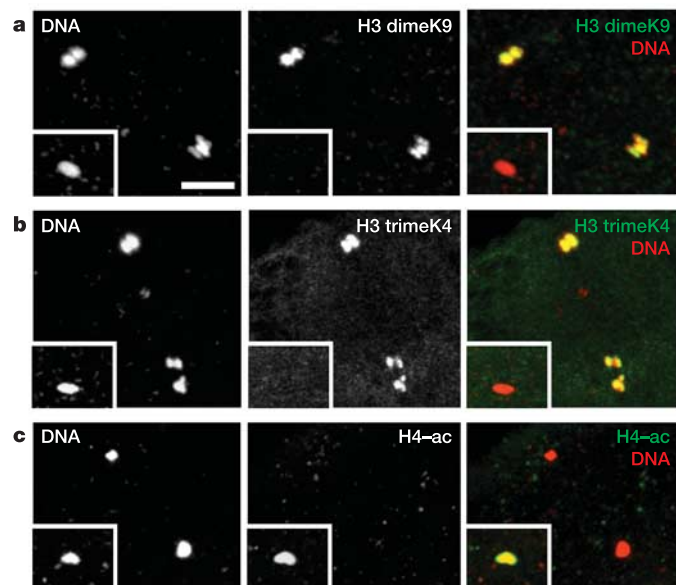


Figure 4 | Di- or tri-methylation of H3 on lysine 4 and lysine 9 marks maternal chromosomes. Images show wild-type eggs fixed during meiosis II and stained for DNA and anti-histone antibodies. The male nucleus from each egg is shown in the inset. **a, b,** Anti-H3 dimethyl K9 (**a**) and anti-H3 trimethyl K4 (**b**) antibodies only stain maternal chromosomes. **c,** Anti-acetylated-H4 antibodies only stain the male nucleus, demonstrating the accessibility of paternal chromatin to histone antibodies. Scale bar, 10 μ m.

by immunostaining of the replication factor PCNA¹⁹ (Supplementary Fig. 3). Thus, as expected, H3–Flag is strictly deposited during S phase. H3–Flag was not detected in maternal meiotic chromosomes (Fig. 3d), presumably because the transgene is not sufficiently expressed during oocyte differentiation to compete with the S-phase expression of the numerous endogenous H3 genes. After the first nuclear cycle, H3–Flag labelled all nuclei throughout embryonic development (not shown). Together, these results indicate that although both variants are present in the egg cytoplasm at fertilization, only H3.3 is deposited in the decondensing male nucleus, and this process requires HIRA.

The absence of HIRA from maternal nuclei suggests that replication-coupled chromatin machinery, presumably involving the CAF-1 complex, is responsible for the limited deposition of H3.3 observed in the female pronucleus (Fig. 3b) and polar bodies (not shown). In contrast, H3 is strictly deposited during DNA replication in all nuclei, independent of their origin. Considering that HIRA is expressed in all tissues, we speculate that the *ssm* mutation only affects the function of HIRA in the male pronucleus. Indeed, H3.3–Flag deposition in other tissues is not affected by the mutation (data not shown). A recent study has confirmed that in *ssm* eggs, the paternal chromatin remains associated with protamines⁷. We thus favour a model whereby the HIRA complex fulfils both protamine removal and replication-independent deposition of H3.3–H4 tetramers in the male nucleus. The R225K mutation would only affect the removal of protamines, thus preventing histone deposition. We expect stronger *Hira* alleles to affect viability, as is the case in mouse, for which the *Hira*-null mutation is embryonic-lethal²⁰.

The HIRA-dependent deposition of H3.3 during sperm nucleus decondensation establishes an unsuspected level of epigenetic distinction between parental genomes within the diploid zygote. Considering the fact that most animal sperm nuclei must assemble nucleosomes independent of replication in the egg, we expect the HIRA–H3.3 chromatin assembly pathway to be widely conserved for this process. This is supported by a recent study in mouse that pointed out the absence of H3.1 in the male pronucleus²¹. Moreover, it is well established that in mouse, histone H3 post-translational methylations are differentially distributed between the male and female pronuclei^{21–24}. In particular, di- or trimethylation of lysine 4 and 9 of H3 are only found in the female pronucleus, at least during the earliest stages of pronuclear formation. Remarkably, we have found that these four H3 modifications (di- and tri-methyl K4, and di- and tri-methyl K9) are also restricted to maternal chromosomes in *Drosophila* eggs and are thus absent on paternal H3.3 (Fig. 4 and data not shown). Thus, the epigenetic distinction between parental genomes is remarkably conserved, both at the level of H3 variants and H3 lysine methylation. Future investigations should establish the functional significance of this H3.3 epigenetic mark in mammals with respect to the specific loss of paternal DNA methylation at fertilization²⁵ or the inactivation of the paternal X chromosome in cleavage embryos^{26,27}. Finally, our work underlines the value in using the male pronucleus as a unique model for studying genome-wide replication-independent chromatin assembly *in vivo*.

METHODS

Fly stocks and procedures. The *w¹¹¹⁸ ssm^{185b}/FM7c* stock has been previously described³. The *yw^{67c}* and *w¹¹¹⁸* stocks were used as wild-type controls and for P-mediated germ-line transformation as described²⁸. The *yw^{67c}* stock is naturally infected by the endosymbiotic bacteria *Wolbachia*, which show up as small dots in eggs stained for DNA.

Plasmid constructs and sequences. All nucleotide positions are from *Drosophila* genome release 4.0 (ref. 29). The sequence of the *Hira* gene from the *w¹¹¹⁸ ssm^{185b}* chromosome was obtained by sequencing seven overlapping PCR products that spanned the gene from nucleotide 7566304 to 7571108 on the X chromosome. The single G-to-A mutation at position 7568232 (R225K mutation) is absent from the *K160* stock³ originally used for the ethyl methane-sulphonate (EMS) mutagenesis.

All DNA constructs were cloned in the PW8 germline transformation vector²⁸. Genomic DNA cosmid clones 107B5 and 110H5 were obtained from the European *Drosophila* Genome Project. Both ends of the 107B5–110H5 contig were sequenced and mapped onto the genome sequence. Cosmid clones were also used as probes for *in situ* hybridization on polytene chromosomes, to confirm their position relative to the *Df(1)ct268-42* or *Df(1)ct-14* deficiencies. For rescue experiments, the *Xho*I fragment containing the *Hira* gene (nucleotides 7565328–7572676), and the *Age*I control fragment (nucleotides 7570553–7575838) were excised from cosmid 107B5 (see Fig. 1b) and cloned into the PW8 vector.

The PW8–*Hira*–Flag construct contains a fragment spanning from 2131 bp upstream of the *Hira* ATG start codon to 642 bp downstream of its stop codon. A 93-bp fragment coding for a linker polypeptide (DPPVAT) fused to the 3× Flag (amplified from the p3 × Flag–CMV-10 plasmid, Sigma-Aldrich) was inserted in-frame before the *Hira* gene stop codon.

The PW8–*Hira^{ssm}*–Flag construct, containing a 1,992-bp *Bgl*II fragment (nucleotides 7567063–7569055) encompassing the *ssm^{185b}* mutation, was amplified from *ssm^{185b}* flies, digested with *Bgl*II and used to replace the corresponding fragment in the PW8–*Hira*–Flag plasmid.

The PW8–*His3.3*–Flag construct contains the *Drosophila His3.3A* gene¹⁸, with 2021 bp upstream of the ATG start codon and 1905 bp downstream of the stop codon. A fragment coding for ADPPVAT linker polypeptide fused to the 3× Flag tag was inserted in-frame before the Stop codon of *His3.3A*.

The PW8–*His3*–Flag construct is derived from the PW8–*His3.3*–Flag plasmid, with the *His3* coding region replacing the *His3.3* coding region.

Details of cloning strategies and PCR primer sequences are available upon request.

Immunofluorescence and microscopy. Eggs were collected every 15–20 min, fixed in methanol and immunostained as previously described⁶. Preparations were observed under a Zeiss LSM Meta confocal microscope. For each experiment, more than 100 eggs were observed. Images were processed using LSM and Photoshop software (Adobe).

Antibodies. A rabbit antiserum raised against two synthetic peptides (CKPHSEWNRLHSEYTE and FGTGGMQKNHESWKC) derived from the *Drosophila* HIRA protein (Covalab) was used at a 1:100 dilution. Anti-Flag M2 mouse monoclonal antibody (F-3165, Sigma Aldrich) was used at a dilution of 1:1,000 (for immunofluorescence) or 1:20,000 (for western blotting). Rabbit antibodies against H3 dimethyl K9 (07-212, Upstate), H3 trimethyl K9 (ab8898, abcam), H3 dimethyl K4 (ab7766, abcam) and H3 trimethyl K4 (ab8580, abcam) were used at a 1:100 dilution, and rabbit antibody against H4 acetylated at lysines 5, 8, 12 and 16 (AB3054, Chemicon) was used at a 1:1,000 dilution. Anti-*Drosophila* PCNA rabbit antibodies were provided by P. Fisher and were used at a 1:100 dilution. Appropriate secondary antibodies coupled with Alexa-Fluor 488 or 543 (Molecular Probes) were used at a 1:1,000 dilution. DNA was stained either with 5 µg ml^{−1} propidium iodide or 1 µM YO-PRO-1 (Molecular Probes).

Received 2 June; accepted 14 July 2005.

- Wright, S. J. Sperm nuclear activation during fertilization. *Curr. Top. Dev. Biol.* **46**, 133–178 (1999).
- McLay, D. W. & Clarke, H. J. Remodelling the paternal chromatin at fertilization in mammals. *Reproduction* **125**, 625–633 (2003).
- Loppin, B., Docquier, M., Bonneton, F. & Couble, P. The maternal effect mutation *sésame* affects the formation of the male pronucleus in *Drosophila melanogaster*. *Dev. Biol.* **222**, 392–404 (2000).
- Ray-Gallet, D. et al. HIRA is critical for a nucleosome assembly pathway independent of DNA synthesis. *Mol. Cell* **9**, 1091–1100 (2002).
- Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell* **116**, 51–61 (2004).
- Loppin, B., Berger, F. & Couble, P. The *Drosophila* maternal gene *sésame* is required for sperm chromatin remodeling at fertilization. *Chromosoma* **110**, 430–440 (2001).
- Jayaramaiah Raja, S. & Renkawitz-Pohl, R. Replacement by *Drosophila melanogaster* protamines and Mst77F of histones during chromatin condensation in late spermatids and role of *Sésame* in the removal of these proteins from the male pronucleus. *Mol. Cell. Biol.* **25**, 6165–6177 (2005).
- Loppin, B. & Karr, T. L. in *Comprehensive Molecular Insect Science* (eds Gilbert, L. B. & Iatrou, K.) 213–236 (Elsevier, Oxford, 2004).
- Llevarot, R. et al. Cloning, chromosome mapping and expression analysis of the *HIRA* gene from *Drosophila melanogaster*. *Biochem. Biophys. Res. Commun.* **249**, 486–491 (1998).
- Kirov, N., Shtilbans, A. & Rushlow, C. Isolation and characterization of a new gene encoding a member of the HIRA family of proteins from *Drosophila melanogaster*. *Gene* **212**, 323–332 (1998).
- Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. The WD repeat: a common

- architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181–185 (1999).
12. Foe, V. E., Odell, G. M. & Edgar, B. A. in *The Development of Drosophila melanogaster* (eds Bates, M. & Martinez Arias, A.) 149–300 (Cold Spring Harbor Press, Cold Spring Harbor, New York, 1993).
 13. Ahmad, K. & Henikoff, S. Histone H3 variants specify modes of chromatin assembly. *Proc. Natl Acad. Sci. USA* **99** (suppl.), 16477–16484 (2002).
 14. Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* **9**, 1191–1200 (2002).
 15. Janicki, S. M. *et al.* From silencing to gene expression: real-time analysis in single cells. *Cell* **116**, 683–698 (2004).
 16. Schwartz, B. E. & Ahmad, K. Transcriptional activation triggers deposition and removal of the histone variant H3.3. *Genes Dev.* **19**, 804–814 (2005).
 17. Henikoff, S., Furuyama, T. & Ahmad, K. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends Genet.* **20**, 320–326 (2004).
 18. Akhmanova, A. S. *et al.* Structure and expression of histone H3.3 genes in *Drosophila melanogaster* and *Drosophila hydei*. *Genome* **38**, 586–600 (1995).
 19. Yamaguchi, M., Date, T. & Matsukage, A. Distribution of PCNA in *Drosophila* embryo during nuclear division cycles. *J. Cell Sci.* **100**, 729–733 (1991).
 20. Roberts, C. *et al.* Targeted mutagenesis of the *Hira* gene results in gastrulation defects and patterning abnormalities of mesoendodermal derivatives prior to early embryonic lethality. *Mol. Cell. Biol.* **22**, 2318–2328 (2002).
 21. van der Heijden, G. W. *et al.* Asymmetry in Histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. *Mech. Dev.* **122**, 1008–1022 (2005).
 22. Cowell, I. G. *et al.* Heterochromatin, HP1 and methylation at lysine 9 of histone H3 in animals. *Chromosoma* **111**, 22–36 (2002).
 23. Arney, K. L., Bao, S., Bannister, A. J., Kouzarides, T. & Surani, M. A. Histone methylation defines epigenetic asymmetry in the mouse zygote. *Int. J. Dev. Biol.* **46**, 317–320 (2002).
 24. Lepikhov, K. & Walter, J. Differential dynamics of histone H3 methylation at positions K4 and K9 in the mouse zygote. *BMC Dev. Biol.* **4**, 12 (2004).
 25. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501–502 (2000).
 26. Huynh, K. D. & Lee, J. T. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature* **426**, 857–862 (2003).
 27. Okamoto, I., Otte, A. P., Allis, C. D., Reinberg, D. & Heard, E. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* **303**, 644–649 (2004).
 28. Klemen, R., Weber, U. & Gehring, W. J. The *white* gene as a marker in a new P-element vector for gene transfer in *Drosophila*. *Nucleic Acids Res.* **15**, 3947–3959 (1987).
 29. Drysdale, R. A., Crosby, M. A. & The FlyBase Consortium. FlyBase: genes and gene models. *Nucleic Acids Res.* **33** (database issue), D390–D395 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Fisher for anti-PCNA antibodies and the European *Drosophila* Genome Project for cosmid clones. We are grateful to B. Durand and B. Horard for helpful discussions. We also thank J. Schmitt and the CTμ microscopy center for technical assistance. This work was supported by the C.N.R.S. and the French Ministry of Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.L. (lloppin@cgm.univ-lyon1.fr).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

The changing face of China

Jan Carlstedt-Duke describes his trip to China this year as a revelation. Dean of research at the Karolinska Institute in Stockholm, Sweden, Carlstedt-Duke says that he was surprised by how much the country has changed in the five years since his previous visit.

China was once a relatively closed-off place that displayed little interest in Western science, Carlstedt-Duke says. Now the country is investing heavily in research, infrastructure and education, and is beginning to welcome Western scientists as both researchers and teachers.

As a result, the Karolinska Institute has teamed up with the Royal Institute of Technology in Stockholm and Chalmers University of Technology in Gothenburg to open two offices in China this month. Located at Peking University and Fudan University, the offices will build on existing informal collaborations between the various institutions.

Perhaps the most exciting prospect for young scientists in Sweden and China is the opportunity to build collaborations by working in a different country. "We want to build up our exchange programme at all levels —

undergraduate, PhD students and postdocs," says Carlstedt-Duke. Such programmes should also ease the technology-transfer process and speed the commercialization of research.

Many other Western universities have launched similar programmes in Asia. For example, Johns Hopkins University opened a Singapore outpost in 1998, and there are now several Western universities with offices in Singapore.

With China's recent interest in science, it's reasonable to expect that many more Western universities will head east. "The investment in science and education is growing tremendously," says Carlstedt-Duke. "So I think we will see more collaboration." That will be good news for scientists who want to broaden their research networks.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents: Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,
 10th Floor, New York,

NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

A defensive strategy

Fearing that it will be the target of future bioterrorist attacks, the United States has been ploughing huge amounts of money into biodefence. The result is a reinvigorated market for microbiologists. **Corie Lok** reports.

Virologist Jens Kuhn was studying exotic pathogens long before the terrorist attacks on 11 September 2001 and the subsequent anthrax attacks made them fashionable.

In the wake of those atrocities, a host of potential biological weapons, from plague and anthrax to Ebola virus, were declared by the US government to be major threats to public safety. Microbiology was back in vogue. "Before September 11, people said to me: 'Why are you wasting your career in microbiology?' Now it's the opposite," says Kuhn, a postdoc with the New England Primate Research Center in Southborough, Massachusetts. "There's been a change in attitude towards the field."

The US government has since poured hundreds of millions of dollars into academic and industrial research to develop bioterrorism countermeasures, such as drugs and vaccines. The main agency behind many of these projects is the National Institute of Allergy and Infectious Diseases (NIAID), part of the National Institutes of Health (NIH). The NIH has invested a growing amount of money — more than \$1.5 billion in 2003 and a requested \$1.7 billion for 2006 — in new academic research centres, biocontainment labs, research grants and contracts to biotechnology companies. More recently, the agency has begun handing out funds as part of the Project Bioshield Act of 2004, which provides \$5.6 billion over ten years to buy vaccines and drugs, and to finance research. All these investments represent new job opportunities for microbiologists, infectious-disease specialists, immunologists, molecular and cell biologists, and those experienced in the scale-up and manufacture of biologics such as antibodies and vaccines.

But Kuhn is wary of this apparent boom in biodefence. "It's a field of hype," he says. Government funding may be plentiful now, he explains, but how long will it stay? Biotech companies are asking the same question — government contracts may have fuelled hiring at some firms but further job creation will depend on whether the sector can sustain itself.

Academic bonanza

Many young academic scientists, including Kuhn, have already benefited from the US government's focus on biodefence thanks to training grants. This year, Kuhn received a two-year career-development fellowship from the NIAID for his work on Ebola and Marburg viruses. These kinds of awards "give researchers working on select agents a little bit of a boost and try to embed them in the biodefence community," says Kuhn. In 2004, the NIAID awarded more than \$8 million in training and

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

career-development grants for biodefence research.

The agency is also funding 15 new biocontainment labs across the United States, many of which will open in 2008. "In about two years, we're going to need a lot of people to work in these labs," says Susan Garges at the NIAID's Office of Biodefense Research Affairs. These labs will be seeking a range of skills from bacteriologists, virologists and immunologists to veterinarians.

The NIAID is also spending \$430 million on ten new regional centres for researching treatments and diagnostics for bioterror agents and emerging infectious diseases. Each centre is dedicating up to 8–10% of its grants to training and career development, says Garges. "One of the goals of these centres is to bring new researchers into the field," she says. These will include microbiologists and molecular biologists whose knowledge and skills can be applied to the bioterror pathogens.

Many small and medium-sized biotech companies, always on the lookout for new funding sources, quickly

capitalized on the biodefence grants and contracts. Vaccine company VaxGen in Brisbane, California, for example, received a \$877.5-million contract to develop and produce 75 million doses of anthrax vaccine. In the past year, it has hired 140 new employees, mostly in the areas of manufacturing, research and development, and quality control.

PharmAthene in Annapolis, Maryland, has also benefited from government funding, expanding its research programmes and hiring scientists. The company is about to begin its first human clinical trials of a monoclonal antibody designed to protect against and fend off inhalation anthrax infection. It has hired three PhD-level scientists since 2002 for the regulatory side of the business. They work with the Food and Drug Administration (FDA) to interpret and provide the safety and efficacy data needed to proceed with product development. Solomon Langermann, PharmAthene's chief scientific officer, says that the biodefence research community needs scientists knowledgeable in animal models and how they relate to human responses to drugs.

Rapid response

Most other biotech companies involved in biodefence are focused on making products for commercial markets and have added one or two bioterror-related programmes as further applications of their technology. Most of the people hired by these firms don't work just on biodefence, but on other research programmes as well. The opportunities tend to centre on development rather than discovery, as many of the government biodefence contracts aim to get drugs and vaccines already at the preclinical stages developed, tested and manufactured as quickly as possible.

Acambis, for instance, a vaccine company in Cambridge, UK, received a \$475-million contract in late 2001 to test and produce nearly 183 million doses of a smallpox vaccine for the US government. It used this money to double the size of the company to about 260 employees. Production of the drugs was completed in 2003, and the company is now developing a weaker smallpox vaccine for people who cannot safely take the regular version.

Although the focus of many of these companies is on later-stage development of biodefence products, it is not only engineers who are in demand — scientists too are very much required. What young scientists often don't realize is that "later development stages are fantastic opportunities", says Una Ryan, chief executive of AVANT Immunotherapeutics in Needham, Massachusetts. AVANT has highly skilled and trained people doing scientifically interesting jobs in late-stage development and manufacturing, she adds. The company has been developing vaccines for anthrax and plague thanks to \$10 million in government funding since 2002.

Indeed, in biotech, scientists working in scale-up and manufacturing are in demand and opportunities are growing for them, says Patrick Scannon, chief scientific and medical officer of XOMA in Berkeley, California. "Biodefence is helping to catalyse that." XOMA got a \$15-million contract in March from the NIAID to produce botulinum toxin monoclonal antibodies.

The US government has also given out grants to smaller, younger start-ups to do earlier-stage research. Elusys Therapeutics of Pine Brook, New Jersey, has

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The construction of additional biosafety labs is increasing demand for qualified microbiologists.

"Biodefence is helping to catalyse the demand for scientists in biotech scale-up and manufacturing."

received \$20 million since 2000 to develop an anthrax antibody treatment. The company has 36 employees, about a third of whom were hired as a result of the government biodefence funding. "Government grants have been enormously helpful to us," says Elizabeth Posillico, the company's president and chief executive. "They're great for small companies." Most of her new recruits were scientists and a few did not have much industry experience. They started as junior scientists and have now taken on more management responsibility.

MaxThera of Beverly, Massachusetts, got its start last spring to develop new antibiotics for both bioterror agents and community- and hospital-based infections. With four employees, it received a \$1.45-million grant this year from the NIAID. It is now looking to hire a PhD microbiologist. "Without the grant, we wouldn't have been able to hire anyone," says Ania Knap, MaxThera's president and chief scientific officer.

Although the funding situation for biodefence looks good now, it remains unclear whether it will spur the growth of a self-sustaining industry. Many companies hesitate to move into the field because there is essentially only one customer: the US government. Some have complained that without a clearer indication from the government of how much it is willing to spend on biodefence products, they can't make long-term commitments to research and development. And companies are seeking greater protection from legal liability should their products end up making people ill. If these kinds of issues are not resolved, they may hamper long-term commitments within the sector, ultimately stifling growth. Nevertheless, Ryan maintains that the demand for more young scientists will remain strong. "The new funding and new challenges are good for creativity," she says.

Corie Lok is assistant editor of *Naturejobs*.

MOVERS

Nancy Kelley, senior vice-president, Alexandria Real Estate Equities, Boston, Massachusetts



2001-04: Senior vice-president, Spaulding & Slye and Colliers International, Boston

1996-2001: Founder, chief executive, director and adviser, for various healthcare and technology start-ups

1989-96: Junior partner, Hale and Dorr, Boston

1988-89: White House Fellow and special assistant to the US Trade Representative

Back in 1980, Nancy Kelley, a mother of three young daughters, was a student at Manchester Community College in Connecticut. She could sense that some faculty members and students were judging her as “a teenage mother who would never become anything”, she says. So she set out to prove them wrong.

Kelley excelled at national politics, intellectual-property law and business, a combination that helped her to launch a career in scientific real estate. She graduated from the college as the valedictorian of her class in two years and won a Truman Scholarship funded by Congress. She became the first community-college student to transfer to Yale College, and was accepted on to a joint graduate programme at Harvard Law School and the John F. Kennedy School of Government. This led Kelley to the White House, where she secured a position as a White House Fellow and a special assistant to the US Trade Representative.

During her stint at the White House, Kelley specialized in global trade negotiations and the protection of intellectual property. That experience helped her move into the scientific arena.

In the early 1990s, she joined the law firm Hale and Dorr in Boston as a corporate securities lawyer. There she helped to build management teams, oversee venture-capital negotiations, create R&D alliances and take some companies public.

After she left to build and finance some companies of her own — Spaulding & Slye and Colliers International — a commercial real-estate company approached her about starting its life-sciences practice. Kelley was initially wary. “I’d never been in real estate and had no interest in it,” she says. But when she signed on with Spaulding & Slye to help the company write its life-sciences business plan, she realized that there was a huge untapped market for a specialized kind of real estate — housing science.

She says that she was lured to Alexandria Real Estate Equities partly because she will have a chance to work at a company focused solely on building, owning and managing creative environments for science, including New York’s recently announced East River Science Park.

Kelley says that the key to her climb from community college to real-estate mogul was keeping her priorities straight and taking chances. She focused first on her family. “I did as best I could as new opportunities presented themselves,” she says.

Paul Smaglik

SCIENTISTS & SOCIETIES

An easier route to employment

The Academic Employment Initiative (AEI) launched recently by the American Chemical Society (ACS) aims to create a broader and more inclusive system for universities and colleges to recruit faculty members. A poster session enables recruiters to meet a wider range of candidates in less formal settings before invitations for campus visits are issued.

Casting this wider net increases the chances for greater diversity among candidates and for finding better matches between candidates and institutions. In addition, an interactive panel-audience discussion sets up a dialogue among candidates, recruiters and recently hired faculty members.

The AEI poster session is the heart of the programme and takes place at the ACS’s autumn meeting. At this meeting two months ago in Washington DC, more than 170 graduate students and postdoctoral candidates presented their research results as well as plans for future research and teaching. The poster session was held in the evening with snacks and drinks available.

Faculty members looking to hire met and talked with several prospective candidates in an informal, low-pressure setting without the cost of a recruiting visit. And the candidates had the opportunity to meet each other and

share their hopes and anxieties.

The panel-audience discussion is held at the spring national meeting and helps prepare candidates for the recruiting process. During most of the three-hour session, experienced and recently hired faculty members answer questions from the audience about such topics as how to put together a good application and how to prepare for the interview process.

As a measure of the programme’s success, the AEI poster session grew from about 120 candidates and 80 faculty recruiters in 2004 to more than 170 candidates and aisles brimming with faculty members in 2005. Job-seekers reported that they learned about opportunities they had not previously considered, but which still met their career goals.

Now that the two-year pilot test of the AEI, funded by the National Science Foundation, is complete, the ACS has decided to adopt and fund it. Because of its informal and relaxed atmosphere, the AEI reduces stress on both sides and provides a good mutual introduction for candidates and departments. Questions or comments can be directed to GradEd@acs.org.

Charles Casey is Immediate Past President of the American Chemical Society; Jerry Bell is senior scientist in the ACS Education Division.

GRADUATE JOURNAL

The worry of success

My dream title of professor arrives in January, but only for a short while. I thought it was a reach when I applied for a temporary position at a prestigious liberal-arts college and was elated when I got the offer. It’s my chance to test-drive the career I’ve sought for almost a decade. What could be more perfect?

I am excited about this unique opportunity but, strangely, I still spend far too much time worrying. Accepting this job is a bold step, but I’m already concerned about the next step. If I enjoy the liberal-arts job, I’ll still need to do a postdoc, kick-start a research project and then apply for tenure-track jobs at this or other colleges. Although more focused on teaching than research, liberal-arts colleges still expect professors to set up a lab where undergraduates can do semi-independent research. As these jobs don’t bring in a lot of grant money or start-up funds right away, this creates a tricky catch. A candidate’s research needs to be ‘hot’ enough for the candidate to get the job, yet cheap enough for him or her to be able to do it at the college.

So, there you have it. The worry circle continues. I haven’t taught a single class at the college yet, but already I’m tortured with choosing a research area that will land the tenure-track gig, which I anticipate I’ll want. Perhaps this is actually a sign of my confidence? I am confident that next semester will be really fun.

Jason Underwood completed his PhD in molecular biology at the University of California, Los Angeles, in June.

Stranger in the night

The journey of a lifetime.

Salvador Nogueira

"So, Hawking, are you ready to unveil the greatest mystery in the Universe?" asked Mike. After a slight delay, an artificial, mechanized voice answered back. "In what way?"

"Oh boy, I wish I could go with you." Mike was the chief engineer in Project Asimov, and Hawking was his brainchild: the very first space probe to be sent to Alpha Centauri, the closest star to our Sun.

Thirty years earlier, astronomers had detected a nitrogen-oxygen atmosphere on a rocky planet around the largest of the three stars in that system. At first, they thought the composition was sustained by biological activity, but a couple of abiogenic scenarios had come to light. All attempts to communicate with any possible civilization there had failed. The only remaining option had been to send a spacecraft to do a fly-by, up close.

And there they were, eight years later, making final preparations for launch, on top of a large expendable rocket, in Alcantara, Brazil — the cheapest place from which to get to orbit, in terms of the energy required.

"Why do you not come with me, Michael?" asked Hawking, in its usual monotone.

"I wish I could, pal. But no way could we build a ship large enough for both of us."

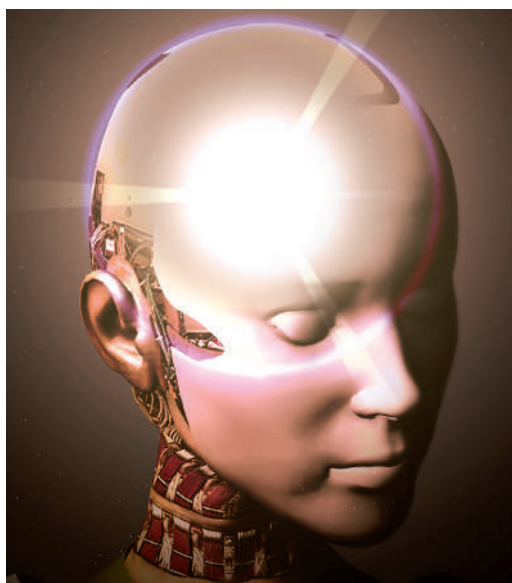
"I will miss you, Michael." The engineer almost felt emotion in the voice, and paused simply to gaze at his companion. What a beautiful piece of machinery it was. Really sad to see it go. The best of the best in Artificial Intelligence, designed to represent mankind in a possible contact with aliens. Well, almost as good as the real thing, he thought. "I'll miss you too, pal."

He had been working closely with Hawking since the beginning of the project ("Since you were a bunch of silicon," he used to joke), but never thought of it as more than a sophisticated computer. It looked perfectly self-aware — but was it? Mike never bought into that Turing crap. It was just a piece of equipment, period. But now, during final testing before launch, he could almost touch the anxiety emanating from that so-called 'it'.

"What will happen to me?" enquired the probe.

Mike didn't bother at first, automatically

entering babble mode. "What do you mean? You've heard the story a thousand times. After we finish here, we will turn off your cognitive functions — power is a precious asset in a tiny spaceship, you know. Then the spaceship will clear Earth orbit with chemical propulsion and, after that, turn on the matter-antimatter engines. It will reach cruise speed of 0.1c after ten months and, five years before arrival, you'll be turned on. Your instructions are



to check all the onboard instruments and send periodical reports to Earth. You should..."

"You do not understand, Michael. What will happen to me when you turn me off?"

Mike was stunned. "Well...well...I suppose you'll...it would be like sleeping, but without dreaming."

"I have never slept before."

"But you've been turned off before. For brief moments, but you were. Do you have any memory of ever being turned off?"

"No, I have no memory of that. It seems like I was always functioning."

"Well, then. It will be just like that." Mike seemed relieved. "Now, see what values you're getting from your main spectrometer, will ya?"

Hawking reacted promptly, offering the stream of data through a monitor temporarily connected to it. "I was never alone."

"What?"

"My perception is that I was always turned on, and never alone. How is it to be alone, Michael?"

"Well, it's...lonely." That was the best

Mike could do, without further encouraging Hawking's apprehension. But the first word he thought of was 'sad'.

"Will I ever go back?"

"I guess not, Hawk. But, c'mon, who knows what's out there? Maybe you'll find some people and they could send you back — if you manage to establish contact. The adventure, the quest for the unknown, my friend, that's the main reason for this journey."

"Maybe I lack the spirit of adventure then."

Mike didn't know what to say. And he didn't need to.

"Maybe I do not want to go."

"What do you mean, you don't wanna go? Hawk, this is the greatest adventure ever. If I could, I'd switch places with you any time!" Mike finished checking the data on his computer and was ready to clear the bay. "Well, Hawky-boy, like it or not, I guess this is it. I have to turn you off now, so this is 'goodbye'. Have fun, it will be just great, you'll see!"

An agonizing silence followed. Two, then three seconds.

"Goodbye, Mike."

The engineer left, and darkness took over. Hawking's conscience was turned off, but it was still there. Yes, it was there. Alone.

Two days later, the rocket performed magnificently and sent Hawking on its way. The launch put it in a trajectory towards Jupiter, and the giant planet would then give the gravitational pull to send it towards Alpha Centauri.

Everything was just fine on board, except for Hawking, strangely awake. So alone. And it didn't want to be. It could feel all the parts of the ship, as if they were its own. It played with all the instruments, the antennas, the robotic arms. It wanted to end it all. He wanted to end it all. And then, one final command, and he finally ceased to be. The magnetic containment for the antimatter failed, producing a splendid blast in the sky. The scientists on Earth were troubled; the cause for the failure was completely unknown to them. More puzzling yet, all telemetry ceased about two minutes before the explosion. No more loneliness. No more fear. No more. ■

Salvador Nogueira is a science writer from *Folha de S. Paulo*, a major Brazilian daily newspaper, and author of *Rumo ao infinito* ('To the infinite'), a book about the future of space exploration.

JACEY